Chapter 3
Sequence Alignment and Homology Modelling

For molecular modeling of proteins in general, the structure of the protein is needed. How can such a structure be obtained? One might consider first a modeling of the protein structure *de novo* or *ab initio* based on the amino acid sequence. There are several approaches described in literature (Fleishman et al. 2006; Yarov-Yarovoy et al. 2006; Taylor et al. 2008; Zhang 2008; Barth et al. 2009; Zaki et al. 2010). For small proteins, these techniques result in suitable structures, which are in good accordance to experimentally derived structures. But it should be taken into account, that with increasing number of amino acids, thus methods are not longer appropriate, because of an exponentially increasing computational time. Thus, other techniques are necessary. One is the technique of homology modelling. This is based on the assumption that proteins of on class have a very similar structure. Thus, if the structure of one protein of a distinct class is evaluated by experimental methods, the structures of all other proteins can be modelled in homology to this experimental template. The technique of homology modelling is used with regard to several GPCRs (Zhang et al. 2006), like the NK1 receptor (Evers et al. 2004), the P2Y$_6$ receptor (Costanzi et al. 2005), the CB2 receptor (Pei et al. 2008), the NKB and NK$_3$ receptor (Ganjiwale et al. 2011), the cholecystokinin-1 receptor (Henin et al. 2006), histamine receptors (Jongejan et al. 2005; Preuss et al. 2007; Jongejan et al. 2008; Lim et al. 2008; Igel et al. 2009; Strasser and Wittmann 2010a; Brunskole et al. 2011) and besides addresses GPCR oligomerization (Simpson et al. 2010).

3.1 Selection of a Template

To be able to start homology modelling, one has to search for an appropriate template structure. A large number of such templates are available at the Protein Data Bank (PDB, http://www.pdb.org). Until end of 2011 a large number of crystal structures were available (Table 3.1). As illustrated by Table 3.1, most crystal structures concern the β_1- and β_2- adrenergic receptor. These crystal structures are of great interest, since different types of ligands, like inverse agonists, antagonists or (partial) agonists are bound. Thus, these crystal structures reveal important information with regard

A. Strasser, H.-J. Wittmann, *Modelling of GPCRs*,
DOI 10.1007/978-94-007-4596-4_3, © Springer Science+Business Media Dordrecht 2013

Table 3.1 pdb-codes of most important crystal structures related to opsin or GPCRs

GPCR	Related pdb-codes
Bovin (rhod)opsin	1F88, 1HZX, 1GZM, 3CAP, 3DQB, 3PQR, 3PXO
Human β_2 adrenergic receptor	2RH1, 2R4R, 2R4S, 3D4S, 3NYA, 3NY8, 3NY9, 3KJ6, 3P0G, 3PDS, 3SN6
Turkey β_1 adrenergic receptor	2VT4, 2YCW, 2YCX, 2YCY, 2YCZ, 2Y00, 2Y01, 2Y02, 2Y03, 2Y04
Human dopamine D_3 receptor	3PBL
Human histamine H_1 receptor	3RZE
Human chemokine $CXCR_4$ receptor	3ODU, 3OE6, 3OE8, 3OE9, 3OE0
Human adenosine A_{2A} receptor	3EML, 2YDO, 2YDV, 3QAK, 3PWH, 3REY, 3RFM

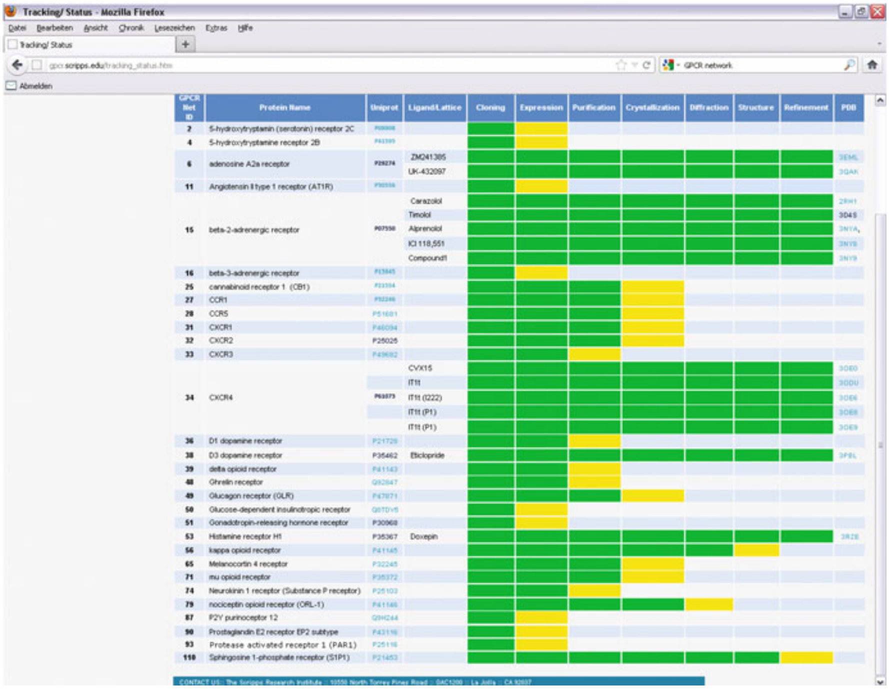
to different conformations of the receptors. Recently, the crystal structure of a ligand bound covalently to the hβ_2R was published (3PDS) (Rosenbaum et al. 2011). Besides the crystal structures of adrenergic receptors, 2010 the crystal structure of the human dopamine D_3 receptor (3PBL) (Chien et al. 2010) and 2011 the crystal strucuture of the human histamine H_1 receptor (3RZE) (Shimamura et al. 2011) was published. In addition to the mentioned crystal structures of biogenic amine receptors, crystal structures of the human chemokine CXCR4 receptor (Wu et al. 2010) and the human adenosine A_{2A} receptor (Jaakola et al. 2008; Lebon et al. 2011; Xu et al. 2011; Dore et al. 2011) are known (Table 3.1).

Thus, if a GPCR has to be modelled an appropriate template has to be chosen. If one likes to model a biogenic amine receptor by homology modelling, the crystal structure of a biogenic amine receptor is suggested to be used as template to solve this task. For modelling of inverse agonists or neutral antagonist in the receptor bound state, a template, representing the inactive conformation should be chosen, whereas a template, representing the active conformation should be used in case of (partial) agonists. Furthermore, the homology between the receptor to be modelled and the template should be as high, as possible. Based on these suggestions, it is the responsibility of the modeller to choose an appropriate template for homology modelling.

Sometimes, a look onto the homepage of GPCR network (http://cmpd.scripps.edu) is very useful. There, you get information about the tracking status of GPCRs which will be crystallized in future (Fig. 3.1).

3.2 Crystal Structures of GPCRs (Source: http://www.pdb.org)

In the appendix, the most important information with regard to all crystal structures of (rhod)opsin or GPCRs is summarized tabular. These tables should give you a fast overview onto available crystal structures, resolution, structure of a cocrystallized ligand, related UniProtKB entries and corresponding literature. Have a careful look onto the section "mutation"! Often, not the wild type receptor is crystallized, instead point mutations were introduced. Thus, if you want to model the receptor, which is crystallized, you may change the amino acids, mutated in the crystal structure, into the corresponding amino acid of the wild type receptor. An overview of the differences in crystal structures is given by the Figs. 3.2–3.6.


Tracking/ Status - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
Tracking/ Status
gpcr.scripps.edu/tracking_status.htm
GPCR network
Abmelden

GPCR Net ID	Protein Name	Uniprot	Ligand/Lattice	Cloning	Expression	Purification	Crystallization	Diffraction	Structure	Refinement	PDB
2	5-hydroxytryptamin (serotonin) receptor 2C	P28335		🟩	🟨						
4	5-hydroxytryptamine receptor 2B	P41595		🟩	🟨						
6	adenosine A2a receptor	P29274	ZM241385	🟩	🟩	🟩	🟩	🟩	🟩	🟩	3EML
			UK-432097	🟩	🟩	🟩	🟩	🟩	🟩	🟩	3QAK
11	Angiotensin II type 1 receptor (AT1R)	P30556		🟩	🟨						
15	beta-2-adrenergic receptor	P07550	Carazolol	🟩	🟩	🟩	🟩	🟩	🟩	🟩	2RH1
			Timolol	🟩	🟩	🟩	🟩	🟩	🟩	🟩	3D4S
			Alprenolol	🟩	🟩	🟩	🟩	🟩	🟩	🟩	3NYA,
			ICI 118,551	🟩	🟩	🟩	🟩	🟩	🟩	🟩	3NY8
			Compound1	🟩	🟩	🟩	🟩	🟩	🟩	🟩	3NY9
16	beta-3-adrenergic receptor	P13945		🟩	🟨						
25	cannabinoid receptor 1 (CB1)	P21554		🟩	🟩	🟩	🟨				
27	CCR1	P32246		🟩	🟩	🟩	🟨				
28	CCR5	P51681		🟩	🟩	🟩	🟨				
31	CXCR1	P46094		🟩	🟩	🟩	🟨				
32	CXCR2	P25026		🟩	🟩	🟩	🟨				
33	CXCR3	P49682		🟩	🟩	🟨					
34	CXCR4	P61073	CVX15	🟩	🟩	🟩	🟩	🟩	🟩	🟩	3OE0
			IT1t	🟩	🟩	🟩	🟩	🟩	🟩	🟩	3ODU
			IT1t (I222)	🟩	🟩	🟩	🟩	🟩	🟩	🟩	3OE6
			IT1t (P1)	🟩	🟩	🟩	🟩	🟩	🟩	🟩	3OE8
			IT1t (P1)	🟩	🟩	🟩	🟩	🟩	🟩	🟩	3OE9
36	D1 dopamine receptor	P21728		🟩	🟩	🟨					
38	D3 dopamine receptor	P35462	Eticlopride	🟩	🟩	🟩	🟩	🟩	🟩	🟩	3PBL
39	delta opioid receptor	P41143		🟩	🟩	🟨					
48	Ghrelin receptor	Q92847		🟩	🟩	🟨					
49	Glucagon receptor (GLR)	P47871		🟩	🟩	🟩	🟨				
50	Glucose-dependent insulinotropic receptor	Q8TDV5		🟩	🟨						
51	Gonadotropin-releasing hormone receptor	P30968		🟩	🟨						
53	Histamine receptor H1	P35367	Doxepin	🟩	🟩	🟩	🟩	🟩	🟩	🟩	3RZE
56	kappa opioid receptor	P41145		🟩	🟩	🟩	🟩	🟩	🟨		
65	Melanocortin 4 receptor	P32245		🟩	🟩	🟩	🟨				
71	mu opioid receptor	P35372		🟩	🟩	🟩	🟨				
74	Neurokinin 1 receptor (Substance P receptor)	P25103		🟩	🟩	🟨					
79	nociceptin opioid receptor (ORL-1)	P41146		🟩	🟩	🟩	🟩	🟨			
87	P2Y purinoceptor 12	Q9H244		🟩	🟨						
90	Prostaglandin E2 receptor EP2 subtype	P43116		🟩	🟨						
93	Protease activated receptor 1 (PAR1)	P25116		🟩	🟨						
110	Sphingosine 1-phosphate receptor (S1P1)	P21453		🟩	🟩	🟩	🟩	🟩	🟩	🟨	

CONTACT US :: The Scripps Research Institute :: 10550 North Torrey Pines Road :: GAC1200 :: La Jolla :: CA 92037

Fig. 3.1 GPCR tracking status. (Status: November 2011; Source: http://gpcr.scripps.edu/tracking_status.htm)

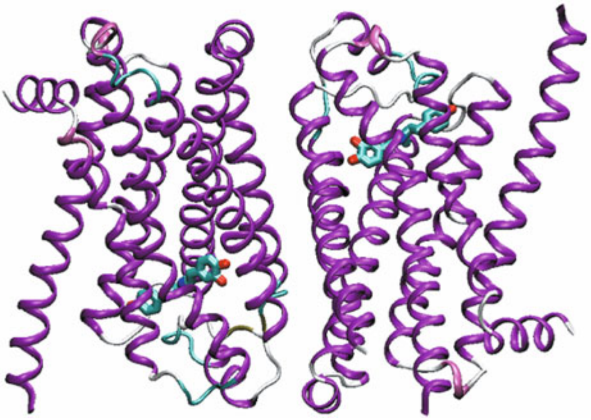


Fig. 3.2 Crystal structure of the turkey β_2R, 2Y00. (Warne et al. 2011)

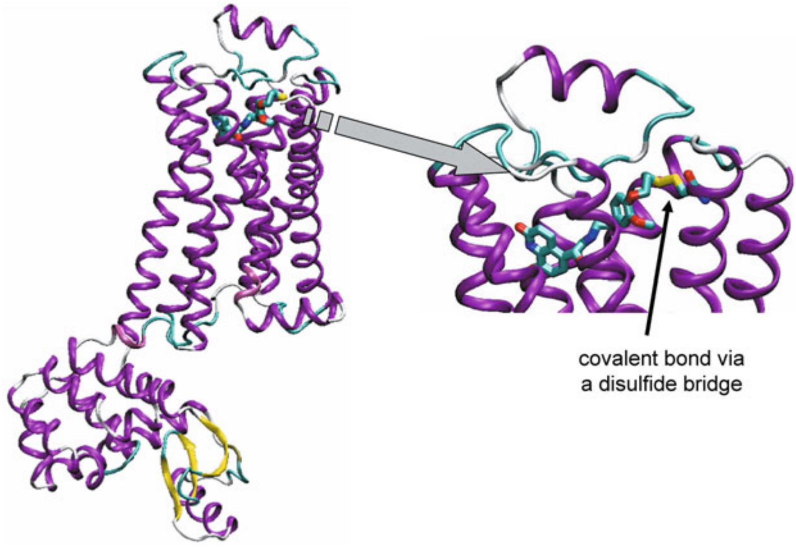
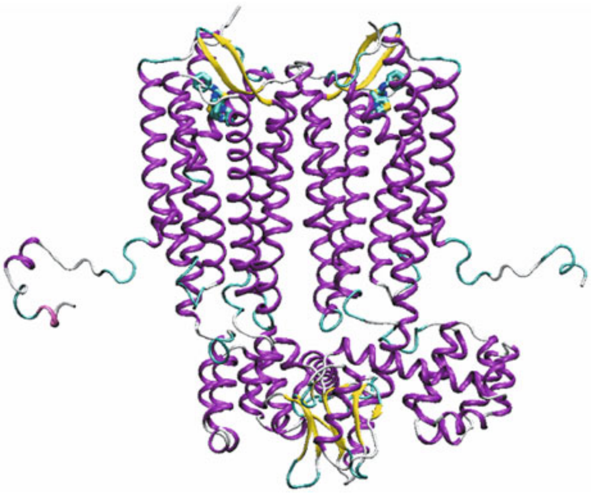


Fig. 3.3 Crystal structure of the human β_2R, 3PDS. (Rosenbaum et al. 2011)

Fig. 3.4 Crystal structure of the human CXCR4, 3ODU. (Wu et al. 2010)



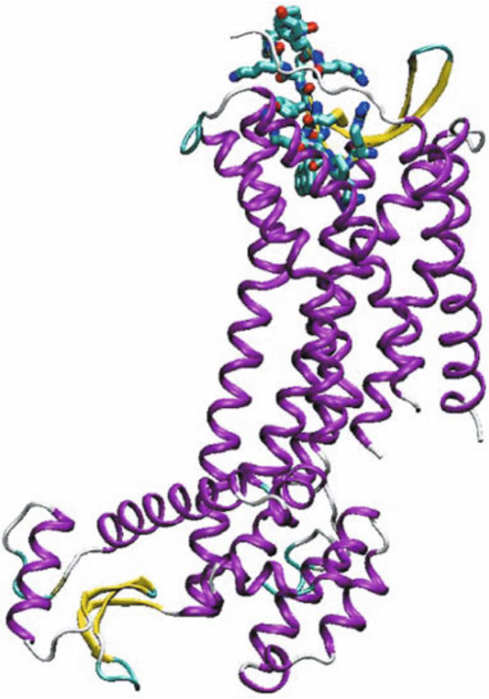


Fig. 3.5 Crystal structure of the human CXCR4, 3OE0. (Wu et al. 2010)

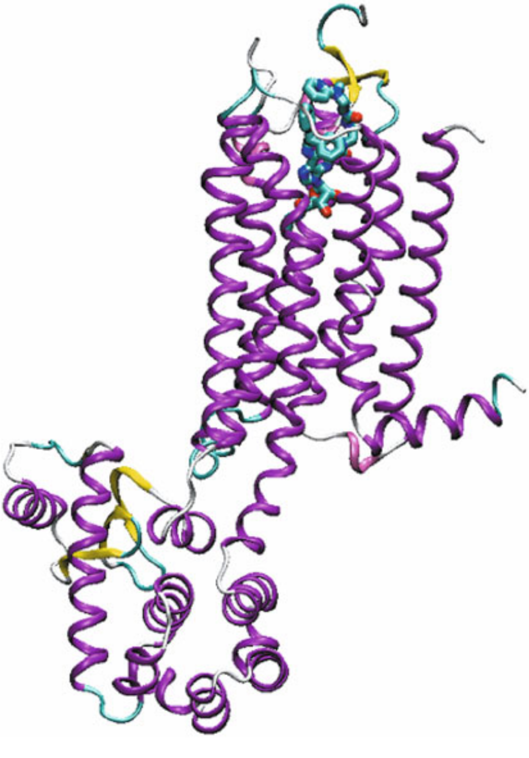


Fig. 3.6 Crystal structure of the human $A_{2A}R$, 3EML. (Jaakola et al. 2008)

3.3 Amino Acid Sequences and Sequence Alignment

Before being able to start the homology modelling, it has to be decided which amino acid of the template sequence corresponds to an amino acid in the target sequence. Therefore, a sequence alignment has to be performed manually or automatically. Clustal (http://www.clustal.org) for example, is a software for multiple sequence alignment. However, before starting with sequence alignment, the corresponding amino acid sequences have to be obtained.

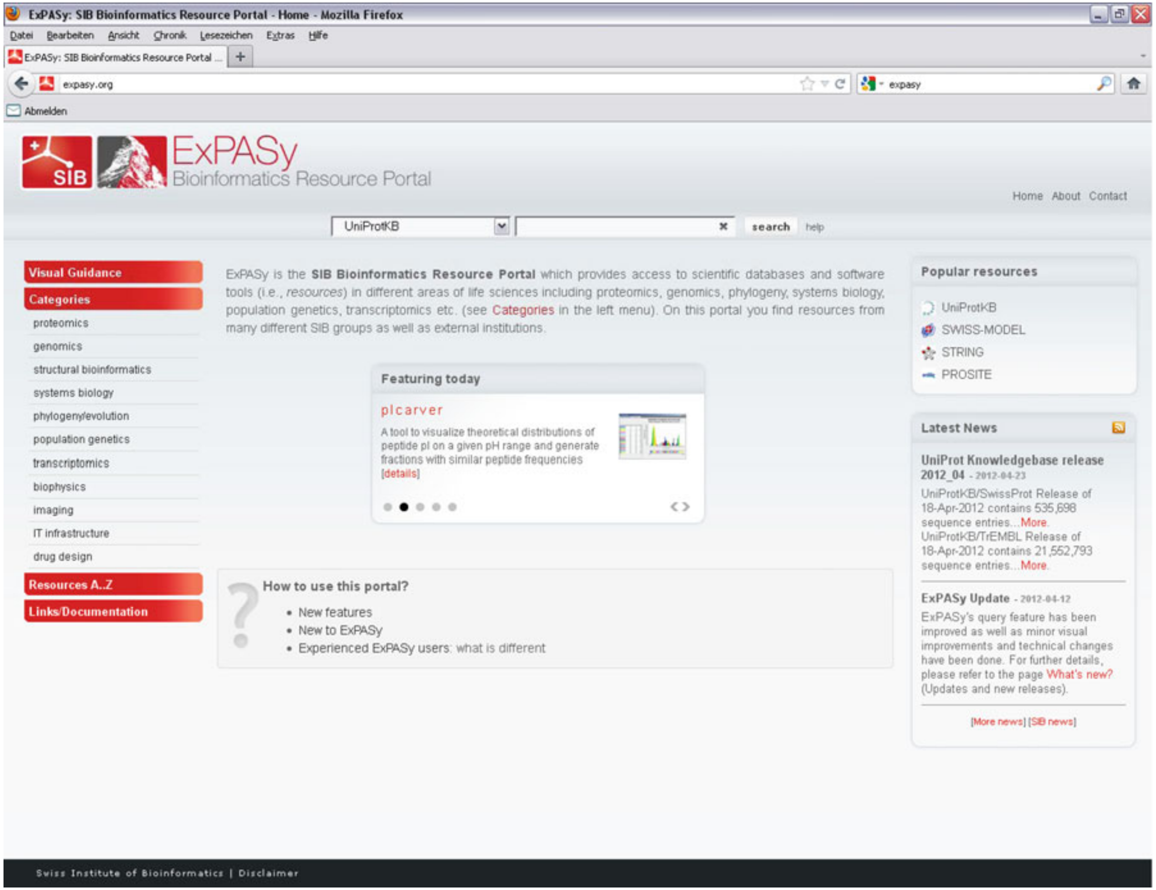
3.3.1 Amino Acid Sequences – Where to Get From?

There are several sources for amino acid sequences present in the internet. One prominent is for example the Expasy Proteomics Server (http://expasy.org) (Fig. 3.7).

Exercise Start your internet browser and open the site http://expasy.org. Now choose "UniProtKB" under the section "query". Then you can type your search string into the field on the right.

Now we want to search for the human adrenergic β_2 receptor. There are different possibilities for the search string. For example, type "adrenergic" and click the "Search" button. Now, more than 900 results, related to "adrenergic" are presented. Scroll, until the receptor of your choice is listed. In our case it is the human adrenergic β_2 receptor with the accession code "P07550". If you want to reduce the number of hits, the search string has to be defined more exactly. Please try "beta adrenergic receptor", "beta-2 adrenergic receptor" and "beta-2 adrenergic receptor human". By defining the search string more exactly, the number of hits can be significantly reduced and it is easier for you to find the hit, you are searching for.

Now, click, onto the corresponding entry with the accession code "P07550" and you get a lot of very useful information about this receptor, including the amino acid sequence. In the section "Regions", the amino acids, related with the N-terminus, C-terminus, intracellular loops, extracellular loops and transmembrane domains are given. This information is very helpful for the sequence alignment later on. In the section "Sequence" you can find the whole amino acid sequence of the protein. For further proceeding on with the amio acid sequence like for sequence alignment, it may be easier for you, to download the amino acid sequence as "fasta" format. To do so, please click onto the string "FASTA". Now you get the amino acid sequence as simple ascii file.



ExPASy is the **SIB Bioinformatics Resource Portal** which provides access to scientific databases and software tools (i.e., *resources*) in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc. (see Categories in the left menu). On this portal you find resources from many different SIB groups as well as external institutions.

plcarver

A tool to visualize theoretical distributions of peptide pI on a given pH range and generate fractions with similar peptide frequencies [details]

How to use this portal?

- New features
- New to ExPASy
- Experienced ExPASy users: what is different

UniProt Knowledgebase release 2012_04 - 2012-04-23

UniProtKB/SwissProt Release of 18-Apr-2012 contains 535,698 sequence entries...More. UniProtKB/TrEMBL Release of 18-Apr-2012 contains 21,552,793 sequence entries...More.

ExPASy Update - 2012-04-12

ExPASy's query feature has been improved as well as minor visual improvements and technical changes have been done. For further details, please refer to the page What's new? (Updates and new releases).

[More news] [SIB news]

Fig. 3.7 Homepage of the expasy server. (http://expasy.org)

Table 3.2 Highly conserved amino acid according to Ballesteros (Ballesteros et al. 2001) of each transmembrane domain of rhodopsin-like GPCRs

TM I	TM II	TM III	TM IV	TM V	TM VI	TM VII
Asn, N	Asp, D	Arg, R	Trp, W	Pro, P	Pro, P	Pro, P

3.3.2 Ballesteros Nomenclature

A careful analysis of the known amino acid sequences of known rhodopsin-like GPCRs by Ballesteros (Ballesteros et al. 2001) exhibited the most conserved amino acid within each of the seven transmembrane domains, which is used as a reference for all other amino acids within the same helix. Within this nomenclature, the term X.YY is used. Therein, X represents the number of the transmembrane domain and YY the position of the residue within the transmembrane domain. The most conserved amino acid within one helix gets the number 50. All other amino acids within the same helix are numbered relative to that highly conserved position 50. The highly conserved amino acids of each transmembrane domain of a GPCR, according to the Ballesteros nomenclature (Ballesteros et al. 2001) are given in Table 3.2.

In Fig. 3.8, the complete amino acid sequence with the conserved amino acids according to Ballesteros (Ballesteros et al. 2001) of the human adrenergic β_2 receptor is presented.

One should pay attention onto the transmembrane regions, as pointed out in Fig. 3.8. As already mentioned the amino acids related to the transmembrane regions are given at http://expasy.org under the corresponding accession code. A comparison to the corresponding crystal structure – if available – shows sometimes differences with regard to the helical region. Let us for example look onto TM III of the human adrenergic β_2 receptor. The transmembrane region is defined from Glu-107 until Val-129 at expasy (Fig. 3.9a). However, a closer look onto the corresponding domain at the crystal structure shows that the helical structure is much longer at both sides (Fig. 3.9b). Thus, the domains are adopted with regard to the amino acid sequence in Fig. 3.9c. Additionally, in Fig. 3.9b, the amio acids Glu-107 and Val-129 are mentioned Glu$^{3.26}$ and Val$^{3.48}$ in the Ballesteros nomenclature. Some additional amino acids are shown in the Ballesteros nomenclature in Fig. 3.9c. For the termini and the loops no corresponding nomenclature is available.

3.3.3 Amino Acid Sequences – Templates

Before performing an amino acid sequence alignment, one has to decide, which structure should be used as template structure for homology modelling. Meanwhile a lot of crystal structures of bovin rhodopsin or GPCRs like the human adrenergic β_2 receptor or turkey adrenergic β_1 receptor are available (see Tab. 3.1 and appendix Important Crystal Structures of GPCRs (Source: http://www.pdb.org)). It cannot be decided overall, which crystal structure should be used as a template for

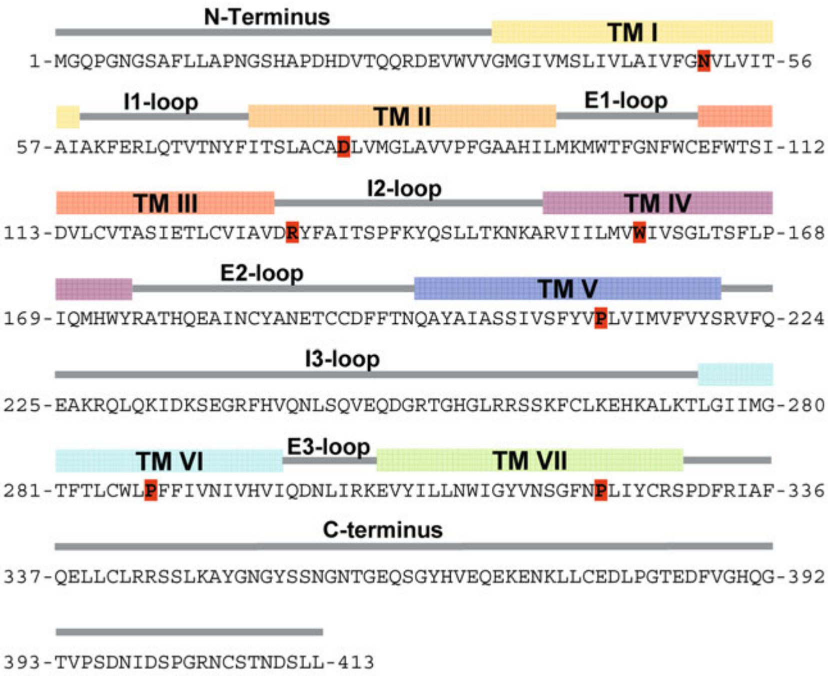

N-Terminus
TM I
1-MGQPGNGSAFLLAPNGSHAPDHDVTQQRDEVWVVGMGIVMSLIVLAIVFGNVLVIT-56
I1-loop
TM II
E1-loop
57-AIAKFERLQTVTNYFITSLACADLVMGLAVVPFGAAHILMKMWTFGNFWCEFWTSI-112
TM III
I2-loop
TM IV
113-DVLCVTASIETLCVIAVDRYFAITSPFKYQSLLTKNKARVIILMVWIVSGLTSFLP-168
E2-loop
TM V
169-IQMHWYRATHQEAINCYANETCCDFFTNQAYAIASSIVSFYVPLVIMVFVYSRVFQ-224
I3-loop
225-EAKRQLQKIDKSEGRFHVQNLSQVEQDGRTGHGLRRSSKFCLKEHKALKTLGIIMG-280
TM VI
E3-loop
TM VII
281-TFTLCWLPFFIVNIVHVIQDNLIRKEVYILLNWIGYVNSGFNPLIYCRSPDFRIAF-336
C-terminus
337-QELLCLRRSSLKAYGNGYSSNGNTGEQSGYHVEQEKENKLLCEDLPGTEDFVGHQG-392
393-TVPSDNIDSPGRNCSTNDSLL-413

Fig. 3.8 Amino acid sequence of the human adrenergic β_2 receptor. The transmembrane domain are presented, as defined at http://expasy.org, accession code P07550. The highly conserved amino acids, defined by Ballesteros (Ballesteros et al. 2001) are marked by *red boxes*

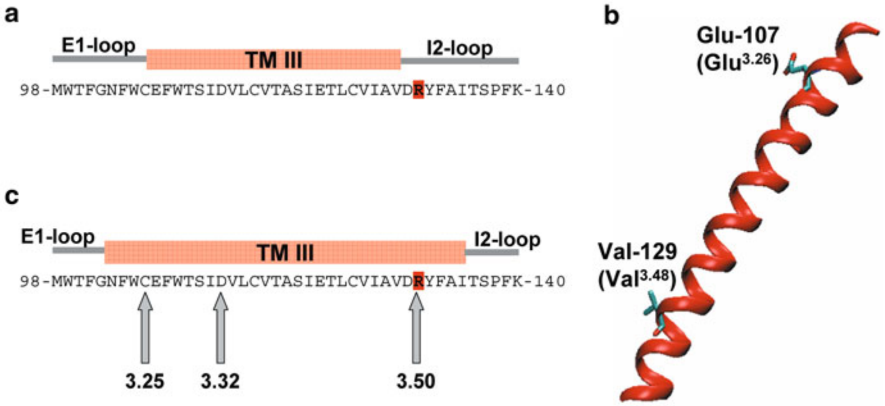

a
E1-loop
TM III
I2-loop
98-MWTFGNFWCEFWTSIDVLCVTASIETLCVIAVDRYFAITSPFK-140
c
E1-loop
TM III
I2-loop
98-MWTFGNFWCEFWTSIDVLCVTASIETLCVIAVDRYFAITSPFK-140
3.25
3.32
3.50
b
Glu-107
(Glu3.26)
Val-129
(Val3.48)

Fig. 3.9 Helical structure of a transmembrane domain. **a** Definition of the TM domain III of the human adrenergic β_2 receptor at expasy (http://www.expasy.org). **b** TM III of the human adrenergic β_2 receptor of a crystal structure. **c** Amino acid sequence of TM domain III, based on the crystal structure

homology modelling. In general, the crystal structure with highest sequence homology to the receptor, which is intended be modelled, should be chosen. Besides that it should be taken into account that different template crystal structures in homology modelling could lead to differences in the resulting homology model. However, the mainly used templates for modelling class A GPCRs are bovine rhodopsin and the human adrenergic β_2 receptor (see appendix Important Crystal Structures of GPCRs (Source: http://www.pdb.org)).

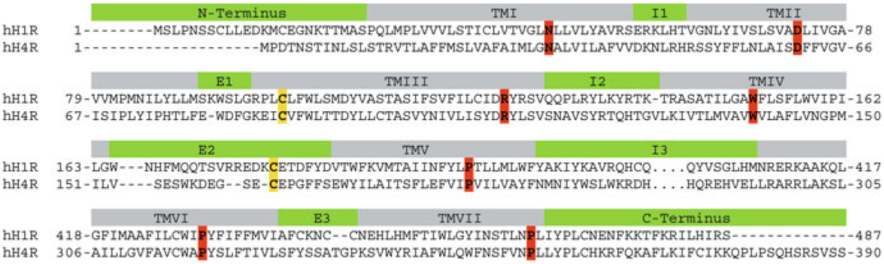
3.3.4 Sequence Alignment

After retrieving the amino acid sequences of the template structure and the destination receptor, the sequence alignment can be performed. There exist several techniques, to perform the sequence alignment. On the one hand, the sequence alignment can be performed manually. The corresponding steps require some time and concentration. On the other hand, there exist several software products, which allow performing an alignment automatically, like clustal (http://www.clustal.org) (see appendix Summary of Important Internet Resources). However, if software is used, it is definitely necessary to check to resulting alignment in order to avoid unexpected mistakes or some inaccuracies.

For a manual sequence alignment, the alignment is performed by several steps:

1. Use the information of the expasy server (http://expasy.org) to get an idea about the amino acids of the seven transmembrane domains for template and target sequence.
2. Perform the sequence alignment for each transmembrane domain in ascending order. Here, it is necessary, that the highly conserved amino acid of each transmembrane domain has the same position in template and target.
3. Now, the alignment for the termini and loops can be performed. There you have to take into account several points:

   - In most crystal structures, the N-terminus and C-terminus are often not complete. Thus, there you can perform the alignment of such regions, but there is no real use in homology modelling, since no template structure is given for such regions.
   - The I1-, E1-, I2- and E3-loop can be aligned easily in most cases to the template sequence. However it should be taken into account, that corresponding loops of different GPCRs could differ in their length. This has to be taken carefully into account later on in the homology modelling. To declare a vacant position in amino acid sequence, a hyphen (-) is used in general.
   - The I3-loop differs significantly in length (from some ten to some hundred amino acids) within the different GPCRs. Additionally, the I3-loop is not completely present in the crystal structures, available up to now. Thus, a complete I3-loop alignment is useless for homology modelling. However, for MD simulations, it will be useful to close the open ends between TM V and TM VI.



```
             N-Terminus                 TMI              I1          TMII
hH1R    1--------MSLPNSSCLLEDKMCEGNKTTMASPQLMPLVVVLSTICLVTVGLNLLVLYAVRSERKLHTVGNLYIVSLSVADLIVGA-78
hH4R    1--------------------MPDTNSTINLSLSTRVTLAFFMSLVAFAIMLGNALVILAFVVDKNLRHRSSYFFLNLAISDFFVGV-66

                   E1              TMIII                 I2              TMIV
hH1R   79-VVMPMNILYLLMSKWSLGRPLCLFWLSMDYVASTASIFSVFILCIDRYRSVQQPLRYLKYRTK-TRASATILGAWFLSFLWVIPI-162
hH4R   67-ISIPLYIPHTLFE-WDFGKEICVFWLTTDYLLCTASVYNIVLISYDRYLSVSNAVSYRTQHTGVLKIVTLMVAVWVLAFLVNGPM-150

                 E2                      TMV                        I3
hH1R  163-LGW---NHFMQQTSVRREDKCETDFYDVTWFKVMTAIINFYLPTLLMLWFYAKIYKAVRQHCQ....QYVSGLHMNRERKAAKQL-417
hH4R  151-ILV----SESWKDEG--SE-CEPGFFSEWYILAITSFLEFVIPVILVAYFNMNIYWSLWKRDH....HQREHVELLRARRLAKSL-305

             TMVI            E3          TMVII                  C-Terminus
hH1R  418-GFIMAAFILCWIPYFIFFMVIAFCKNC--CNEHLHMFTIWLGYINSTLNPLIYPLCNENFKKTFKRILHIRS--------------487
hH4R  306-AILLGVFAVCWAPYSLFTIVLSFYSSATGPKSVWYRIAFWLQWFNSFVNPLLYPLCHKRFQKAFLKIFCIKKQPLPSQHSRSVSS-390
```

Fig. 3.10 Manual alignment of the hH$_4$R to the hH$_1$R. *green*: termini and loops; *grey*: transmembrane domains; *red boxes*: highly conserved amino acids (Ballesteros et al. 2001); *yellow*: highly conserved cysteine, establishing a disulfide bridge to the upper part of TM III; -: missing amino acids; the amino acids of the I3-loop are not shown completely, which is indicated by *dots*

Therefore, some amino acids of the beginning and end of the I3-loop are modelled correctly and the gap is closed by an alanine chain.

- The E2-loop has to be aligned very carefully. It has to be taken into account, that there is a highly conserved disulfide bridge between the E2-loop and the upper part of TM III. Thus, the corresponding cysteine has to be positioned correctly.

An example for an alignment of the human histamine H$_4$ receptor to the human histamine H$_1$ receptor is shown in Fig. 3.10.

3.4 Homology Modelling

3.4.1 *Modelling of the Transmembrane Domains*

The helical transmembrane domains can be easily modelled straight forward. Therefore, only the amino acid side chains have to be changed into the side chain of the destination with appropriate modelling software.

3.4.2 *Modelling of Loops*

In general the transmembrane domains of different GPCRs consist of the same number of amino acids. Thus, the homology modelling of transmembrane domains is quite easy and can be performed straight forward. In case of intra- or extracellular loops, which are connecting the transmembrane domains, differences in number of amino acids of a loop between different GPCRs can occur. This is the case for the E2- or E3-loop between hH$_1$R and hH$_4$R (Fig. 3.10). Small gaps can be closed with "loop search" modules by using appropriate software. For some biogenic amine

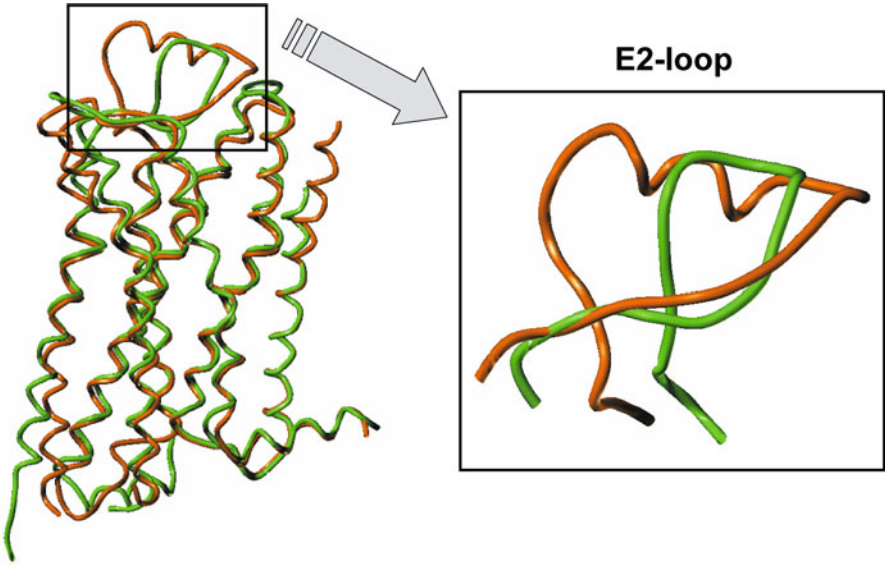


Fig. 3.11 Different conformations of the E2-loop based on crystal structures

receptors, an influence of extracellular loops, especially the E2-loop, onto the binding of ligands to the receptor was shown (Lim et al. 2008; Brunskole et al. 2011). Thus, a correct modelling of the loops is very important. Most of the loops are resolved by crystal structures. However, this is often not the case with regard to the extracellular loop E2 and this is not the case with regard to the intracellular loop I3.

Since the E2-loop is in contact with the binding pocket, the E2-loop has to be modelled completely. If you look onto different crystal structures with complete E2-loop, you can see different conformations (Fig. 3.11).

Thus, you have to decide carefully, which template is to be used for modelling of the E2-loop. A large number of crystal structures are obtainable for the human adrenergic hβ_2 receptor. But the hβ_2R is a special case: There are two disulfide bridges in the E2-loop (Fig. 3.12), whereas in most others GPCRs there is only one disulfide bridge in the E2-loop, connecting the E2-loop with the upper part of the TM III.

A part of the E2-loop of the hβ_2R exhibits a helical structure, but this is not the case for all other GPCRs. Thus, you have to decide carefully, if it would be appropriate to use two different template structures for homology modelling: one for the E2-loop and one for the remaining parts of the receptor. However, the E2-loops are widely different in their length, thus, in most cases, the E2-loop cannot be modelled by changing an amino acid side chain of the template into the side chain of the destination. Thus, you have to use also techniques, like "loop search". For only one loop search, the number of amino acids is too long, and you would get bad results. Thus, it is better, to use at least one fixed point. This is the highly conserved cysteine, connecting the E2-loop by a disulfide bridge with the upper part of TM III (Fig. 3.10).

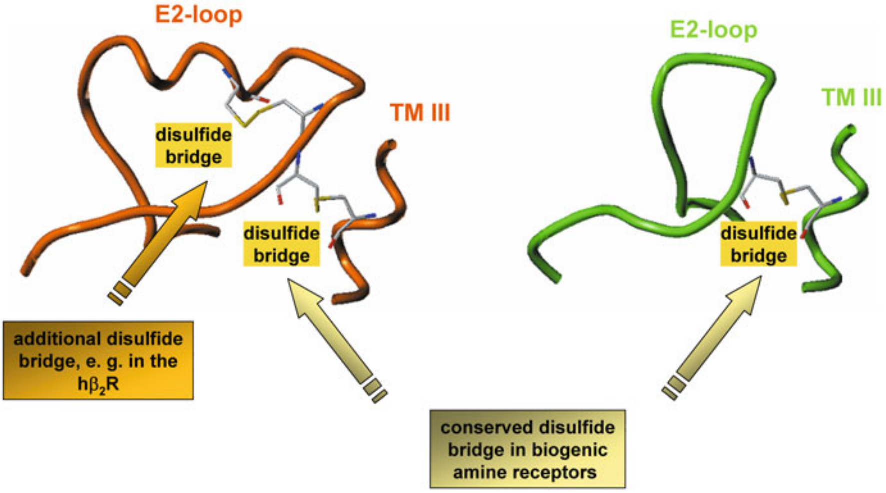

E2-loop
TM III
disulfide bridge
disulfide bridge
additional disulfide bridge, e. g. in the hβ$_2$R
E2-loop
TM III
disulfide bridge
conserved disulfide bridge in biogenic amine receptors

Fig. 3.12 Different numbers of disulfide bridges in the E2-loop

3.4.3 Modelling of Internal Water

A detailed analysis of the crystal structures of GPCRs reveals that there are internal, highly conserved water molecules present (Fig. 3.13). Several studies showed that these water molecules are involved in the hydrogen bonding within the receptor. Based on the published data, it can be suggested, that these water molecules are essential for stabilizing the receptor or important for receptor activation (Pardo et al. 2007). Thus, in order to generate a stable receptor model, the water molecules which are localized/crystallized within the receptor should be included into the homology model.

3.4.4 Modelling of the C-Terminal Part of the Gα Subunit or the Whole Gα Subunit

Based on several studies it is suggested, that a GPCR in its active conformation interacts in the intracellular part with the Gα subunit. There is only small knowledge about the receptor – G protein interaction. However, recently, the crystal structure of opsin, cocrystallized with eleven amino acids of the C-terminus of the Gα subunit (Scheerer et al. 2008) and a complete GPCR – G protein complex (Rasmussen et al. 2011) were published. A detailed analysis of the corresponding crystal structures (3DQB, 3SN6) shows, that the C-terminus of the Gα subunit is deeply bound in a pocket between the transmembrane domains. Leaving out this part of the Gα will result in some problems in subsequent molecular dynamic simulations. In general, if molecular dynamic simulations of a receptor are performed, the receptor is embedded in its natural surrounding. Thus, if the C-terminal part of Gα or the whole Gα is

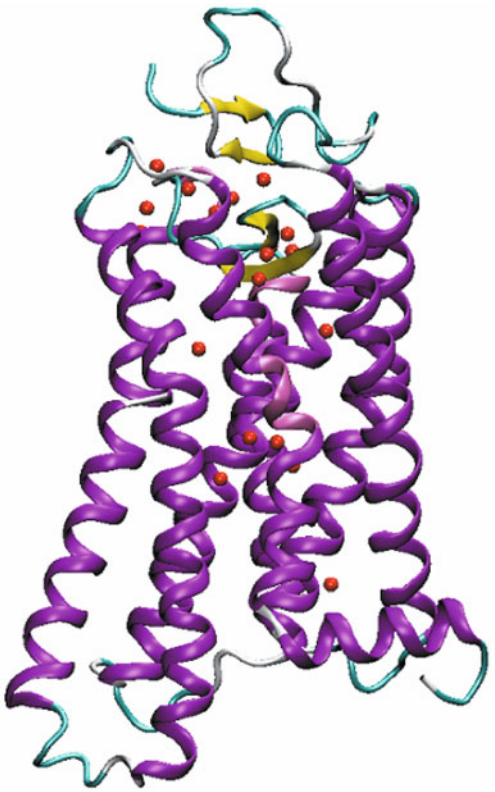


Fig. 3.13 Crystal structure of bovin rhodopsin (1GZM) with internal water (*red balls*). (Li et al. 2004)

missing, the resulting free space is filled with water molecules. Water molecules are highly polar and thus have completely other (surface) properties than the C-terminal part of Gα. Thus, leaving out the C-terminal part of Gα and substitution by water molecules in molecular dynamics can lead to instabilities of the receptor during the molecular dynamic simulation. Thus, it is suggested, to include at the whole Gα or least the C-terminal part of Gα in a homology model. Be aware, that each GPCR couples to a distinct Gα subunit (Fig. 2.8).

3.4.5 Refinement of the Receptor Model

After finishing the homology modelling, several checks of the complete model should be performed. A typical error of beginners in molecular modelling is presented in Figs. 3.14–3.16. During homology modelling, some amino acid side chains have to be mutated into the correct amino acid side chain. Sometimes, especially with regard to long side chains or aromatic rings, collisions between the side chains arise. There are two types of collisions: In the first type, two side chains are in close contact, as shown in Fig. 3.14. In most of these cases, energy minimization is sufficient to

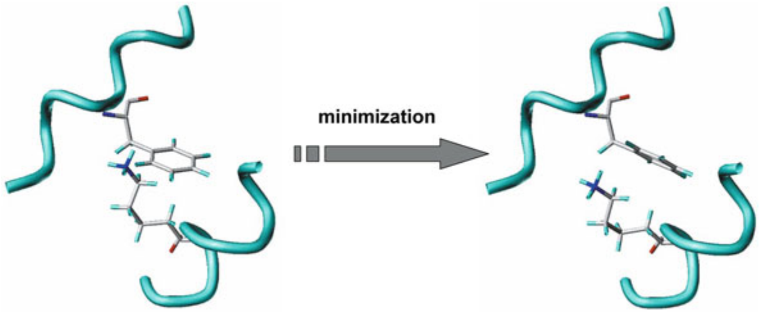
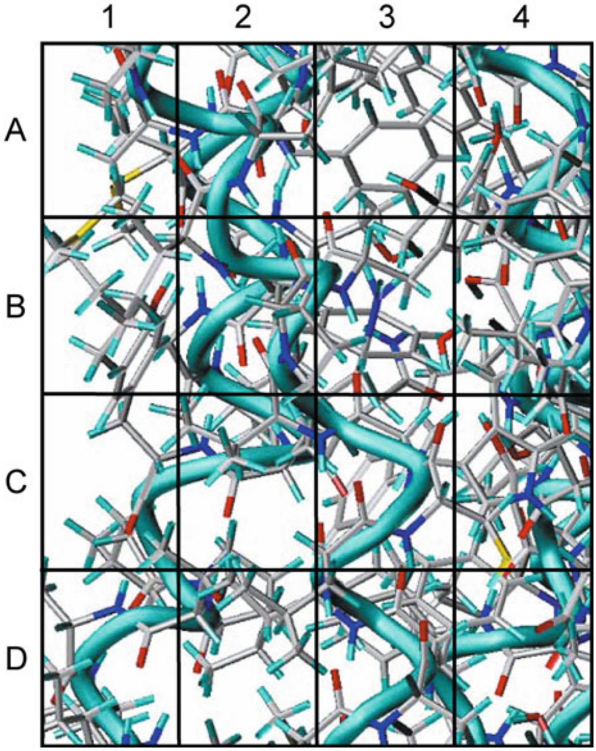


Fig. 3.14 Close contact between the atoms of a lysine and phenylalanine: *Left*: before minimization, *right*: after minimization

Fig. 3.15 Part of a protein structure after minimization. What is the problem?



remove the collision and suitable structures might be obtained. The second type of collision is a more difficult pitfall, which is illustrated in Figs. 3.15 and 3.16. Look carefully onto the Fig. 3.15. Where is the problem?

After a careful look onto the picture you may see, that there is a problem with regard to a lysine and phenylalanine in box B3. This is also illustrated in Fig. 3.16.

Here, a long amino acid side chain, like present in lysine, is located within an aromatic ring, like present in tyrosine, phenylalanine, tryptophane or histidine.

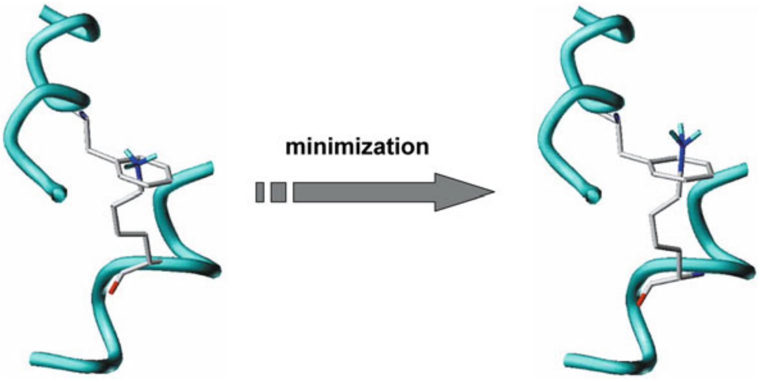


Fig. 3.16 Wrong close contact between the atoms of a lysine and phenylalanine: *Left*: before minimization, *right*: wrong structure after minimization

Unfortunately, a large number of modelling software minimizes a protein, containing such type of wrong structure. And additionally in most cases the potential energy is negative. Thus, one might conclude that all is well. However, often during molecular dynamic simulation, problems occur and the simulation stops with an error. If this is the case, you have to go back to your starting structure and look for the error. Often, an error, similar to that described above (Fig. 3.16) causes the problem. A similar problem can occur not only within the protein, but also between protein and lipid molecules. If there are collisions between amino acid side chains, one has to decide, how to remove this collisions. In general, there are two possibilities: First, one can simply perform an energy minimization. But in some cases, this could lead to artefacts, especially, if two aromatic moieties are linked together. Thus, it is suggested, that one looks separately onto each collision and tries to remove the collisions by carefully changing the corresponding dihedral angles.

After completing these steps, the homology model can be energetically minimized. Here it is suggested, that the energy minimization is performed step by step. In order to avoid structural artefacts, induced by minimization, it is important, that the backbone of the transmembrane domains is provided with position restraints during a first minimization. In a subsequent minimization steps, the receptor can be minimized without any position restraints. Afterwards, the model should be checked, addressing the following items and if everything is correct, one can start with further modelling studies, like docking or molecular dynamic simulations.

□ Check for the correct amino acid sequence

□ Check for the presence of the disulfide bridge between the E2-loop and the upper part of TM III

□ Check for the correct configuration of the amino acids

□ Check for collisions or bad contacts between amino acid side chains