TOWARDS SOLVING THE DOPAMINE G PROTEIN COUPLED RECEPTOR MODELLING PROBLEM

A thesis submitted to the University of Glasgow

for the degree of

Doctor of Philosophy

in the Faculty of Science

by

Christopher D. Wood, B.Sc. (Hons.), M.Sc.

Department Of Chemistry

University Of Glasgow

1st December, 1994

CONTENTS

LIST OF FIGURES

LIST OF TABLES

DECLARATION

This thesis is less than 100,000 words in length, exclusive of tables, figures, bibliographies and appendices. The work reported in this thesis was performed entirely by myself unless otherwise cited or acknowledged. The contents have not been previously submitted for any other degree.

signed

21st December, 1994.

This Ph.D. is dedicated to little Lowri Seren Jones - arrived 2 months early and has cried ever since. The Lord Bless You And Keep You.

ACKNOWLEDGEMENTS

I would like to express special thanks to my supervisor Professor Neil Isaacs - without him this Ph.D. would not have been completed. I must also acknowledge the support I have received from Drs Kelvin Tyler, Paul Mallinson, Paul Emsley, Adrian Lapthorn, Steve Prince, Jeremy Beauchamp, Peter Holliman and Paul Newman; also Gerry McDermott, David Gourley, John Maclean, Calum Sutherland and David Adam. I would also like to thank Dr. Christopher Gilmore for introducing me to the Cambridge Structural Database and giving me the opportunity to annually demonstrate this important scientific resource to Glasgow's third year Chemistry undergraduates - a most stimulating and rewarding experience. Special thanks also to Stuart MacKay and Christopher Edwards for lots of help with UNIX fixes; Noel Ruddock for help with FORTRAN programming techniques; Organon (UK) Ltd. for two years full funding; Professor Geoff Webb (Head of Department) for essential help in arranging additional finance in my final year; all the staff at Central Computer Services (John Buchanan, Lorna Hughes, Elizabeth McAlavey, Dr. Derek Higgins, Jem Taylor, Dr. Ian Walker, Paul Rosenberg and Ian Woodrow); Dr. Keith Vass for tuition in the use of the powerful GCG Sequence Analysis Package and user support; Leigh Blue Caldwell, David M. F. Brewster, Iain C. McInnes and Andrew M. Scott - who lifted my spirits on long nights in the basement cluster room. Dr. Rao for lifting my spirits on long days in the cluster room. I would also like to thank and acknowledge Professor Hilary Evans (Director of Biomolecular Sciences at Liverpool John Moores University, UK) for past support and encouragement and likewise Dr. Florian Müller-Plath (formerly of the S.E.R.C Daresbury Laboratory, Theory and Computational Science Division). Betty Forbes (Departmental Secretary - now retired) for lots of encouragement particularly when I needed it. Glasgow University Accommodation Office Staff for helping me out with accommodation very close to the department when I went down with the nasty *Sarcoid* - in particular: lovely Francis! Hugh McLenaghan for being a constant source of distraction! Lastly, I would like to thank my family and friends back home in Cardiff (Caerdydd) and Barry, South Wales - their constant support and encouragement helped a great deal.

Abbreviations

2D	Two diminsional
3D	Three diminsional
βAR	β-adrenergic receptor
β-ARK	β-adrenergic receptor kinase
bRh	Bacteriorhodopsin
d	Residues per turn of α-helix
D_n	Dopamine D_n subtype receptor; $n \in \{1, 2 \ldots 5\}$
DA	Dopamine - a natural catecholamine ligand.
DAR	Dopamine receptor
DARs	Dopamine receptors
e_n	Extracellular loop n; $n \in \{1, 2, 3\}$
EM	Energy Minimisation
EMBL	European Molecular Biology Laboratory - main laboratory is located in Heidelberg, Germany
fsec	Femtosecond (10^{-15} s)
GC	Grand Challenge
GEZ	Goldman, Engelman and Steize or GES average hydrophobicity curve
GOR	Garnier, Osguthorpe and Robson (GOR) method of secondary structure prediction
GPCR	G protein coupled receptor
GPCRs	G protein coupled receptors
GCG	Graphics Computer Group - Winsconsin Sequence Analysis Software Package for the UNIX and VMS systems
GFlop	Giga Flop - one thousand million floating point operations per second
GRKs	G protein coupled receptor kinases
H-bonding	Hydrogen bonding
HSSP	Database of Homology-derived Structures and Sequences and Sequence alignments of Proteins
IH	Internal Homology Scan Algorithm
i_n	Intracellular loop n; $n \in \{I, II, III\}$

i_n	Intracellular loop n; $n \in \{1, 2, 3\}$ - alternative nomenclature
KD	Kyte-Doolittle Plot
MB	Mega-byte - million bytes
MD	Molecular Dynamics
MFlop	Mega Flop - one million floating point operations per second
MHz	Mega hertz - million cycles per second
mips	Million instructions per second
MPP	Massive parallel processing
MPTP	1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine
MSA	Multiple Sequence Alignment
MSAs	Multiple Sequence Alignments
MSDOS	Microsoft Disc Operating System™
MSF	Multiple Sequence File
PDB	Protein Database at Brookhaven
PHD	Profile network prediction Heidelberg
PRRs	Proline rich regions.
ps	Picosecond (10^{-12} s)
ptmh	putative transmembrane helix
ptms	putative transmembrane segment
Q_3	Three states of protein conformation: helix, sheet and coil
Q_4	Four states of protein conformation: helix, sheet, coil and turns
QSAR	Quantative structure-activity relationship analysis
prc	Photosynthetic Reaction Centre
RAM	Random access memory
RC	Photosynthetic Reaction Centre
rH	Rhodopsin
RH	Reverse Homology Scan Algorithm
RHM	Rigid helix motions
SAR	Structure-activity relationship analysis
SDS	Sodium dodecylsulfate
TM	Transmembrane
wt	wild type

The IUPAC-IUB one-letter amino acid code has been used in the PHD secondary structure prediction algorithm results, the GCG generated Multiple-Sequence-File. Else-where, the three letter amino acid code has been used (in brackets):

A	(Ala)	Alanine
C	(Cys)	Cysteine
D	(Asp)	Aspartate
E	(Glu)	Glutamate
F	(Phe)	Phenylalanine
G	(Gly)	Glycine
H	(His)	Histidine
I	(Ile)	Isoleucine
K	(Lys)	Lysine
L	(Leu)	Leucine
M	(Met)	Methionine
N	(Asn)	Asparagine
P	(Pro)	Proline
Q	(Gln)	Glutamine
R	(Arg)	Arginine
S	(Ser)	Serine
T	(Thr)	Threonine
V	(Val)	Valine
W	(Trp)	Tryptophan
Y	(Tyr)	Tyrosine

Abstract

The overall aim of this work has been to furnish a model of the dopamine (DA) receptor D_2. There are currently two sub-groups within the DA family of G protein coupled receptors (GPCRs): D_1 sub-group (includes D_1 and D_5) and the D_2 sub-group (includes D_2, D_3 and D_4). Organon (UK) Ltd. supplied a disk containing the PDB atomic co-ordinates of the integral membrane protein bacteriorhodopsin (bRh; Henderson *et al.*, 1975 and 1990) to use as a template to model D_2 - the aim being to generate a model of D_2 by simply mutating the side-residues of bRh. The assumption being that bRh had homology with members of the supergene class of GPCRs. However, using the GCG Wisconsin GAP algorithm (Devereux *et al.*, 1984) no significant homology was detected between the primary structures of any member of the DA family of GPCRs and bRh. However, given the original brief to carry out homology modelling using bRh as a template (see appendix 1) I felt obliged to carry out further alignments using a shuffling technique and a standard statistical test to check for significant structural homology. The results clearly showed that there is no significant structural homology, on the basis of sequence similarity, between bRh and any member of the DA family of GPCRs. Indeed, the statistical analysis clearly demonstrated that while there is significant structural homology between every catecholamine binding GPCR, there is no structural homology what so ever between any catecholamine binding GPCR and bRh.

Hydropathy analysis is frequently used to identify the location of putative transmembrane segments. However, is difficult to predict the end positions of each ptms. To this end a novel alignment algorithm (DH Scan) was coded to exploit transparallel supercomputer technology to provide a basis for identifying likely helix end points and to pinpoint areas of local homology between GPCRs. DH Scan clearly demonstrated characteristic transmembrane homology between different subtype DA GPCRs. Two further homology algorithms were coded (IH Scan and RH Scan) which provided evidence of internal homology. In particular IH Scan independently revealed a repeat region in the 3rd intracellular loop (iIII) of D_4 and RH Scan revealed palindromic like short stretches of amino acids which were found to be particularly well represented in predicted α-helices in each DA receptor subtype. In addition, the profile network prediction algorithm (PHD;

Rost *et al.*, 1994) predicted a short α-helix at greater than 80% probablility at each end of the third intracellular loop and between the carboxy terminal end of transmembrane VII and a conserved Cys residue in the forth intracellular loop.

Fourier analysis of catecholamine binding GPCR primary structures in the form of a multiple-sequence file suggested that the consensus view that only those residues facing the protein interior are conserved is not entirely correct. In particular, transmembrane helices II and III do not exhibit residue conservancy characteristic of an amphipathic helix. It is proposed that these two helices undergo a form of helix interface shear to assist agonist binding to a Asp residue on helix II. This data in combination with information from a number of papers concerning helix shear interface mechanism and molecular dynamic studies of proline containing α-helices suggested a physically plausible binding mechanism for agonists.

While it was evident that homology modelling could not be scientifically justified, the *combinatorial approach* to protein modelling might be successfully applied to the transmembrane region of the D_2 receptor. The probable arrangement of helices in the transmembrane region of GPCRs (Baldwin, 1993) which was based on a careful analysis of a low resolution projection map of rhodopsin (Gebhard *et al.*, 1993) was used as a guide to model the transmembrane region of D_2. The backbone torsion angles of a helix with a middle Pro residue (Sankararamakrishnan *et al.*, 1991) was used to model transmembrane helix V. Dopamine was successfully docked to the putative binding pocket of D_2. Using this model as a template, models of D_3 and D_4 were produced. A separate model of D_1 was then produced and this in turn was used as a template to model D_5.

1. The Dopamine Receptors And Their Medical Significance

1.1 Background

Neurological conditions involving psychomotor disorders and some psychoses are influenced by the activity of dopaminergic neurones and by drugs that interact with neuronal dopamine (DA) receptors (Niznik *et al.*, 1992). In particular, rigidity in Parkinson's disease, hallucinations in schizophrenia and Alzheimer's disease, dyskinesia in Huntingtons's chorea, and spontaneous oral dysinesia in the elderly (Seeman, 1987; Seeman *et al.*, 1987; Seeman and Niznik, 1990). These receptors may also play a crucial role in drug addiction and alcoholism (Blum et al., 1990), opiate-withdrawal syndrome (Schulteis and Koob, 1994; Harris and Aston-Jones, 1994) and Tourette's syndrome (Civelli *et al.*, 1992). In the case of Parkinson's disease, too little dopamine causes rigidity, so clinical treatments use agonists as drugs. In contrast, schizophrenia is caused by too much dopamine and so treatments largely rely on prescribing neuroleptics (i.e. antagonists or partial agonists) to block the receptors to prevent hallucinations. The ultimate aim of this work is to provide a template (or series of templates) in the form of a theoretical receptor model with defined pharmacophoric points in space. Such a template could prove to be of immense value in aiding Medicinal Chemists to model new drugs particularly for the treatment of psychomotor disorders and psychoses.

1.2 The Classification/Nomenclature of the Dopamine Receptors

On the basis of pharmacological, biochemical, and physiological criteria, receptors for dopamine have been classified into two types, termed DAe and DAi on the basis of the excitatory or inhibitory properties of DA (reviewed by Horn, 1990). Various nomenclatures have caused a considerable amount of confusion, but DA receptor classification has been clarified by the gene cloning of the DA receptors subtypes: D_1, D_2, D_3, D_4 and D_5 (Dearry, *et al.*, 1990; Dal-Toso *et al.*, 1989; Sokoloff *et al.*, 1990; Van Tol *et al.*, 1991 and Sunahara *et al.*, 1991 respectively). It is quite possible that new members

will be cloned. Though the classification of the dopamine family of receptors is subject to some controversy, it is universally agreed that D_1, D_2, D_3, D_4 and D_5 are all members of a large gene family of hormone/neurotransmitter receptors that exert their biological actions via signal transduction pathways that involve guanine nucleotide-binding proteins. Hence, the term G-protein coupled receptors (GPCRs) is frequently used to describe them. While D_2 (together with D_3 and D_4) dopamine receptors inhibit the activation of adenyl cyclase and appear to couple to numerous other effector systems, D_1 (together with D_5) dopamine receptors, stimulate adenyl cyclase and subsequently activate cAMP-dependent protein kinases (Niznik, 1987).

Up until the close of the 1960s, there was no need to postulate the existence of more than one type of DA receptor. However, the picture became quite complicated onwards from the 1970s. Several groups suggested the existence of several types of DA receptors on the basis of various pharmaceutical findings. The classical DA receptors (i.e. those that can be selectively stimulated with apomorphine and inhibited with haloperidol) was challenged by Cools and van Rossum (1976). They noted that mammalian DA receptors could be selectively stimulated by 3,4-dihydroxyphenylimino-2-imidazolidine (DPI) and inhibited by ergometrine. Consequently, they postulated two principle types of mammalian DA receptors which can be divided into excitation-mediating (DA_e or D-1 activity) and inhibition-mediating (DA_i or D-2 activity) receptors[1]. The situation became more complicated with the dominance of ligand binding as an important tool for research. Examination of the binding of various radioactive ligands of the DA receptor led to a nomenclature involving four possible binding sites, i.e. D_1, D_2, D_3 and D_4 (Seeman, 1980). This classification was clarified by the finding that the D_3 and D_4 receptors were high affinity states of D_1 and D_2 respectively (Urwyler and Markstein, 1986).

The cloning of five distinct genes has helped to further clarify the nomenclature (Niznik and Van Tol, 1992). Sibley and Monsma (1992) proposed that a hierarchical system of nomenclature be applied to the dopamine receptor family. They put forward the proposal

[1] This classification scheme fell on rocky ground since the key compound 3,4-dihydroxyphenylimino-2-imidazolidine (DPI), which they claimed was a selective DAi agonist (i.e. bound only to DA_i) turned out to be a mixed α_1/α_2-adrenoceptor agonist (van Oene and Horn, 1985).

that this nomenclature should be based on both structural and pharmacological criteria. Where the transmembrane (TM) sequence homology is greater than 50% compared with a previously cloned subtype, then it should be designated as a member of that subfamily using an A, B, C etc. nomenclature - all things being equal on the pharmacological front. Hence, D_1 and D_5 would be called D_{1A} and D_{1B} respectively; D_2, D_3 and D_4 would be called D_{2A}, D_{2B} and D_{2C} respectively. They noted that D_1 sub-family has nanomolar affinity for the antagonist compound SCH23390 while the D_2 sub-family has picomolar to nanomolar affinity for the antagonist spiperone. They proposed that if a new DA receptor is cloned it could be called D_{3A} only if binding affinity to SCH23390 or spiperone did not match that characteristic of the D_1 or D_2 sub-families.

However, the author perceives several potential problems with this classification method as it takes no account of the overall topology of any newly cloned DA receptor. For example, the D_1 sub-family has a characteristic hydropathy plot which easily distinguishes it from that of the D_2 sub-family[2]. What if the next newly cloned DA receptor has nanomolar affinity for SCH23390, a TM homology with the D_1 sub-family of 51% but has a hydropathy plot which is clearly different from D_{1A} or D_{1B}? According to Sibley and Monsma's proposed classification scheme, this newly cloned DA receptor would be called D_{1C} when clearly it would represent a completely new DA receptor sub-family. To make matters more confused, the current version of the popular SWISSPROT data-base (release 26.0 7/93) has opted for a compromise: D_1 and D_5 are referred to as DADR and DBDR (hence some similarity here to Sibley and Monsma's preferred names: D_{A1} and D_{A2} respectively). However, the remainder of the DA receptors are called: D2DR, D3DR and D4DR. The author has adopted the following scheme: dopamine receptors are simply referred to in the order in which they were cloned: D_1, D_2, D_3, D_4 and D_5. Members of the D_1 subfamily of dopamine GPCRs are: D_1 and D_5; members of the D_2 subfamily are: D_2, D_3 and D_4 ; by default, D_2 signifies the long isoform and D_{2S} the short isoform of D_2.

There are two isoforms of the D_2 receptor: short and long - both are identical except for an insertion of 29 amino acids in the third intracytoplasmic loop which arises from two

[2] Hydropathy analysis suggests that the D_1 sub-family clearly has a shorter hydrophilic 3rd intracellular loop and a longer hydrophilic carboxy tail than members of the D_2 sub-family.

mRNAs produced from the same gene by alternative splicing. Fujiwara *et al.* (1991) refer to the "short" and "long" forms as: D2(415) and D2(444). Malmberg *et al.* (1993) observed that several antipsychotic compounds (including clozapine and substituted benzamides) bound with 2 to 3 fold higher affinity to the D2B (short) than the D2A (long) isoforms of D_2. Malmberg's use of "D2A" and "D2B" terms completely clashes with Sibley and Monsma's (1992) proposed classification scheme. Hayes *et al.* (1992) also use the same terms as Malmberg's: D2A and D2B for the long and short isoforms of D_2 and so is also in complete disagreement with Monsma's hierarchical scheme. Adding to the confusion in the nomenclature, numerous authors (e.g. Flitz, *et al.*, 1993; Montmayeur *et al.*, 1993; Leysen *et al.*, 1994) use the terms D2L and D2S to refer to the long and short isoforms of D_2. Jose *et al.* (1992) use the terms: D2short and D2long. Andersen *et. al.* (1990) did suggest that where two forms of the same receptor arise by alternative splicing of an exon in the same gene, then the two forms should be referred to as isoreceptors of that particular type. Using Monsma's hierarchical scheme in combination with Andersen's suggestion the short and long isoforms of D_2 could be named: D2A(415) and D2A(444) respectively - though D2A(short) or D2A(long) would also be satisfactory.

DA receptors located only in the CNS are referred to as centrally acting DA receptors[3]. Peripheral DA receptors (which were thought to be identical to the centrally acting DA receptors) are also found. As Jose *et al.* (1992) point out, these are generally referred to as: DA1 or DA2 subtypes. The DA1 receptors approximate the pharmacological profiles of the D1 and D5 receptors, whereas those of the D2A subtype, roughly approximate those of the D2 receptors. However, Jose *et al.* (1992) have identified a receptor with unique characteristics and have named it the DA2k receptor (discussed in more detail in the next section). The *ad hoc* way in which these GPCRs are named merely adds even more weight to the growing consensus that a clear scheme of nomenclature/classification is long overdue - a scheme which should allow for possible cloning of new dopamine receptor subtypes.

De-Keyser (1993) noted that all subtypes of dopamine receptors that have been identified so far still fit in the traditional D1/D2 dopamine receptor classification scheme. However,

[3] The classification of the dopamine receptors with particular reference to their neuroanatomical distribution is reviewed by Brucke *et al.*, 1991.

De-Keyser then goes onto propose a hierarchical system of nomenclature ignoring Sibley and Monsma's (1992) proposals for a hierarchical system of nomenclature for the DA family of receptors. As if to underline the basic premise that confusion is the norm in classifying the DA receptors, De-Keyser comments that "the current status of the different subtypes of the D1 and D2 receptor families in human brain remains unclear."

Gene cloning has a very important role in clarifying the classification of the dopamine family of GPCRs (and GPCRs in general) and improving our understanding of the role played by these receptors (briefly reviewed by Andersen *et al.*, 1990). For example, in 1991 there was a growing feeling that classical neuroleptics (neuroleptics that induce extrapyramidal side effects) and atypical neuroleptics (neuroleptics that do not induce these side effects) may work at least partially through the dopamine D_1 receptors whereas classical neuroleptics are generally considered to work via the dopamine D_2 receptors. Indeed, Ellenbroeck *et al.* (1991) has produced data that suggests haloperidol brings about its therapeutic (and extrapyramidal side effects) via blockade of the dopamine D_2 receptors, whereas clozapine produces its therapeutic effects (with minimum extrapyramidal side effects) via blockade of the D_1 receptors. However, the cloning and characterisation of D_4 (Van Tol *et al.*, 1991; also see: Van Tol *et al.* 1992 and Shaikh *et al.*, 1993) reveals it has a higher affinity for clozapine than any other dopamine receptor. Since D_4 is classified as a D_2 type receptor (D_{2C} - using Monsma's hierarchical classification scheme) this obviously suggests that the conclusions of Ellenbroeck *et al.* (1991) are wrong simply because they were not aware of the existence of the D_4 receptor subtype.

Recent sequence analysis work by Donnelly *et al.* (1994) has concluded that the classification of GPCRs into subclasses based upon the nature of their ligands is over-simplified. Sequence similarity dendograms indicated that the evolution of GPCRs occurs at two sites: the G protein-coupling regions and the ligand binding site and this results in convergent as well as divergent evolution. Hence, D_4 is an example of a D_2 type dopamine with a ligand binding site that is close to the D_1 binding site. Also, the findings of Donnelly *et al.* (1994) suggest that it is also important not to classify GPCRs on the basis of TM homology and ligand binding as proposed by Sibley and Monsma (1992). In chapter five (table 5.3.5) random shuffling of whole sequences to calculate Z scores clearly

differentiates between the members of D_1 and D_2 dopamine GPCRs. In addition, the length of the third intracellular loop and carboxy tail are important (reviewed in chapter 3). Members of the D_1 sub-family have a short third intracellular loop and a long carboxy tail; in contrast, Members of the D_2 sub-family have a longer third intracellular loop and a much shorter carboxy tail.

To sum up, there are two distinct sub-families within the dopamine family of GPCRs. The D_1 sub-family includes two dopamine receptor sub-types: D_1 and D_5. The D_2 sub-family includes three dopamine receptor sub-types: D_2, D_3 and D_4. D_2 sub-type dopamine receptors inhibit the activation of adenyl cyclase and appear to couple to numerous other effector systems, have a long third intracellular loop and a short carboxy tail. D_1 sub-type dopamine receptors, stimulate adenyl cyclase and subsequently activate cAMP-dependent protein kinases, have a short third intracellular loop and a much longer carboxy tail. Calculation of Z scores using whole sequences can be used to differentiate between D_1 and D_2 sub-type dopamine receptors.

1.3 Brain Disorders - Financial Costs to Nations

Given the role of the dopamine receptors in disorders of a neurobiological nature it is important to consider the tremendous costs involved, both economic and social costs. A recent report produced for the National Foundation (discussed in a *Nature editorial*, 4th June, 1992) for Brain Research in Washington, DC, suggests that brain disorders cost the US $401,000 million per year:

	$ (millions)
Psychiatric disease:	136,000
Alcohol abuse:	90,000
Drug abuse:	71,000
Neurological disorders:	104,000
Total:	401,000

This figure represents 7.3% of US gross domestic product (GDP), though the validity of these cost estimates have been challenged (Coyle, 1992).

1.3.1 Personal Costs of Parkinson's Disease and Schizophrenia

There is a huge cost (social as well as financial in terms of lost earnings) to the sufferer and immediate members of their family. Frequently, close relatives have to give up work to look after a loved one or are forced to perform double duties with consequent risks to their health (West, 1991). To illustrate this point here are two extracts:

- Manifestations of Parkinson's Disease (West, 1991): cardinal manifestations: tremor, rigidity, bradykinesia (slowness of movement) and postural instability (chasing one's own centre of gravity). Secondary manifestations: incoordination, micrographia (writing difficulties), blurred vision, impaired upgaze, blepharospasm (spasm of the eyelids), dysarthia (slurring of the speech), dysphagia reflex (difficulty in swallowing), sialorrhea (drooling), masked faces (expressionless face), monotone voice, hand and foot deformities, festinating gait (short, quick, tottering steps; appearing to be constantly falling forwards), cogwheel rigidity (muscle relaxes and stiffens intermittently giving a jerky movement), dystonia (muscle spasm), edema (swelling of the extremities), kyphosis (curvature of the spine), pain and sensory symptoms, constipation, urinary urgency, hesitancy and frequency, loss of libido, impotence, freezing, dementia, depression.

- Schizophrenia - as experienced by a past student of the author: "I see things and hear voices. The voices are not invited and I have nothing to do with them. They come out of the blue and are very disturbing. They have told me to put my head in a box. The medicines make me sick and flatten my mind. I deny that I am sick. Sometimes I don't take them (the drugs) but this causes me to hear more things and I upset my family. If I miss them I can be very sick. No one knows if I will get better - I was fine up until after I graduated with a science degree from Kingston Poly - went to South Africa and got involved in the politics there. Probably a big mistake, my mind just seemed to go away and I returned to Cardiff. I do silly things - physically shake friends or people I don't really

know very well. The illness makes me want to walk a lot and I walk around the same places constantly. I walk around the training centre opening doors and closing them again for no reason at all. I guess it is something to do with coping with the illness and possibly my reaction to the drugs."

Hence, there is a clear need to develop accurate 3D models of these receptors to aid the medicinal chemists in the rational drug design approach.

1.3.2 Epidemiology of Parkinson's disease

The diagnosis of Parkinson's disease is difficult because there is no objective test (e.g. diagnostic test based on the existence of a genetic marker). Consequently, diagnosis depends on clinical judgement. Parkinson's disease is estimated to affect approximately 100,000 people in the UK (West, 1991). It is the second most common neurological condition (dementia is the first) to affect the elderly British population. But the condition is found all around the world and affects white and black peoples, though there is some disagreement about whether white peoples are more prone to the disease. It ranks equally with stroke in causing disability. Studies have shown that the rate of new cases varies between 16 to 21 per 100,000 per year. The incidence of the disease increases until it peaks at 75 years of age and then declines. The US Bureau of the Census estimates that the 65 and onwards age group in the USA will rise from 25 million (1980) to 32 million by the year 2000. Similarly, UK population projections bases on mid-1987 population data, predicts the 65-74 age group will grow from 5.0 million (1988) to 6.8 million by the year 2013. Hence, the rate of clinical diagnosis for this disease is expected to grow together with the increasing age of the population. In addition, 10% of new cases occur in the under 40s age group.

There is currently no research indicating a genetic basis for acquiring the disease. Whether environmental factors play a decisive role is also not clear. The chance discovery that 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP, Coleman *et al.*, 1988) is a neurotoxin which destroys the dopamine producing cells of the primate brain, reinforced the hypothesis

that Parkinson's disease is due to contact with neurotoxic agent(s) in the environment. However, the evidence overall neither favours the protagonists for the environmental theory or the genetic theory.

1.3.3 Epidemiology of Schizophrenia

Schizophrenia is a devastating affliction with bizarre symptoms. Half a million people in the UK, and two million in the US, will suffer from the disease at some point in their lives (Understanding the inner voices; *New Scientist,* 9th July, 1994). Schizophrenia has *positive* and *negative* aspects. Early on in the disease the hearing of inner voices and high frequency of hallucinations (positive aspects of schizophrenia) tend to give way to withdrawal and difficulty with holding even simple conversations with others (the negative aspects of the disease).

Crow and Harrington (1994) point out that psychotic illnesses (schizophrenia and schizoaffective and affective psychosis) have a lifetime prevalence of 2-3% and probably occur at a similar rate in all human societies. This suggests a genetic disease independent of environmental triggers such as viruses (e.g. prenatal exposure to influenza - Crow, 1994). Indeed, Crow points out in the New Scientist article that no etiologically significant environmental precipitants have been identified, and this suggests that these diseases are primarily genetic. Given that schizophrenics are less likely to have children than the general population Crow argues that the gene responsible for schizophrenia must carry advantages to help conserve it in the gene pool. Also, that episodes of illness can be ameliorated by dopamine neuroleptics (i.e. antagonists) targeted particularly at D_2.

2. Dopamine

2.1 Dopamine

Dopamine (DA; figure 2.1) was first synthesised in 1910, but its pharmacological properties were not realised due to the fact that its sympathomimetic actions are weaker than those of adrenaline and noradrenalin (Horn, 1990) - figure 2.1. Nearly four decades later (1938), it was shown that DA occurred in human urine and in 1939 Blaschko first proposed that L-DOPA (figure 2.1) and DA were intermediates in the biosynthesis of noradrenalin and adrenaline from L-tyrosine (Horn *et al.*, 1979). The first direct evidence that DA played a vital role in the functioning of the mammalian brain came in the late 1950s. For example, Carlson *et al.* (1957) demonstrated that L-DOPA was able to reverse the action of a sedative (reserpine) in mammals. DA was found in significant concentrations in the neostriatum and patients afflicted with Parkinson's disease were found to have little if any DA in their corpus striatum. This finding led to the administration of L-DOPA to successfully treat rigidity in Parkinson's disease - L-DOPA is able to cross the blood-brain barrier and is decarboxylated in the brain tissue to DA.

2.2 Molecular Properties.

Since DA is the natural ligand we must ask the basic question: *what information can we gleam from simply looking at the molecular structure of the DA molecule?* Knowledge of the DA structure and its properties would help us determine the likely structure and properties of the binding pocket. Combined with a multiple sequence alignment (Feng and Doolittle, 1987) of the catecholamine binding GPCRs (DA, α and β adrenergic receptors) would help identify key amino acid residues in the receptor binding site.

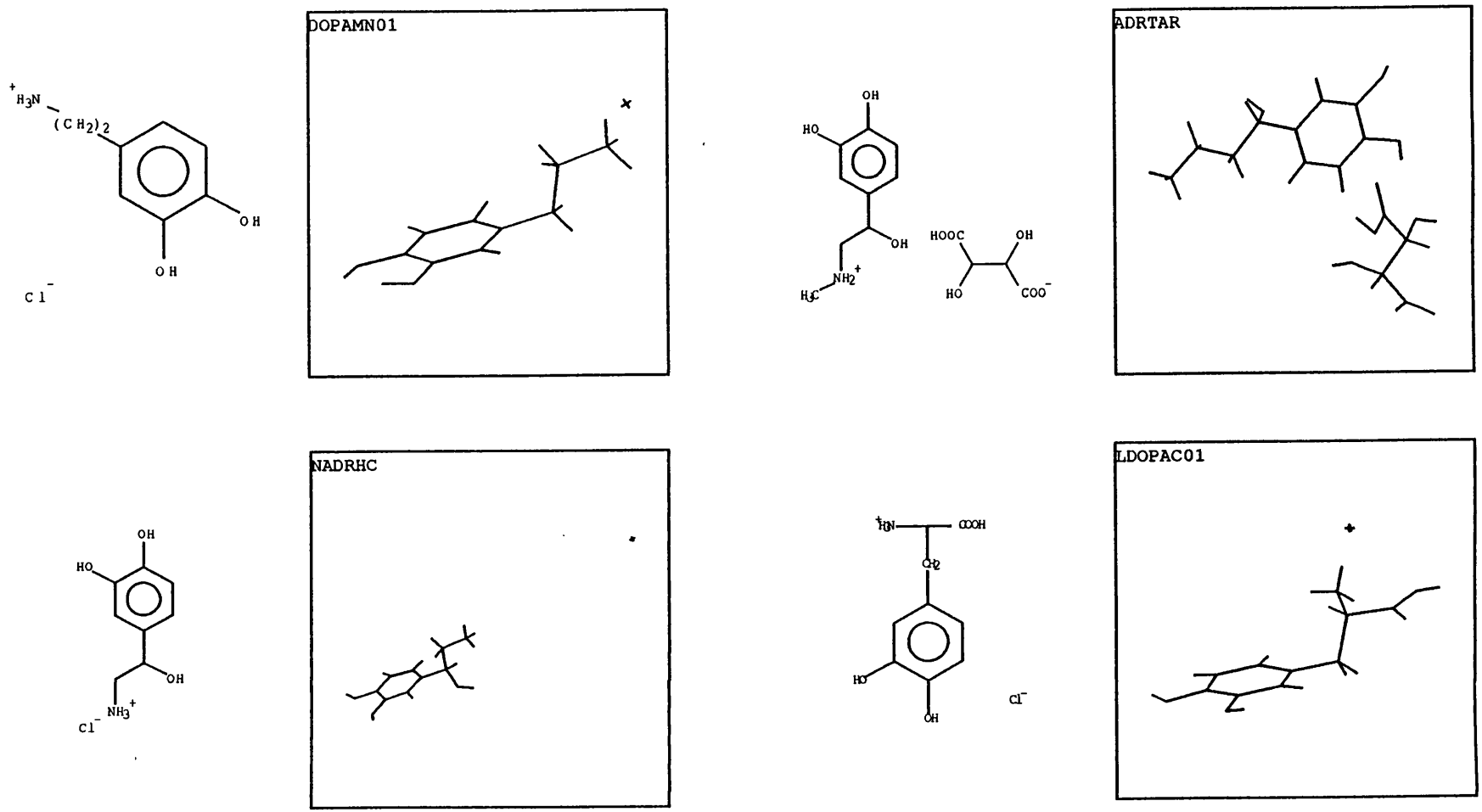


Figure 2.1 Molecular structures of catechol amine ligands extracted from the Cambridge Structural Database (Allen and Kennard, 1983). Top left: dopamine [dopamine hydrochloride; Giesecke, 1980; entry: DOPAMN01]; Top right: adrenaline [(-_-adrenaline hydrogen (+)-tartrate; Carlstrom, 1973; entry: ADRTAR]; Bottom left: noradrenaline [noradrenaline hydrochloride; Carlstrom and Bergin, 1967; entry NADRHC]; Bottom right: L-DOPA [L-DOPA hydrochloride; Mostad and Romming, 1974; entry: LDOPAC01]

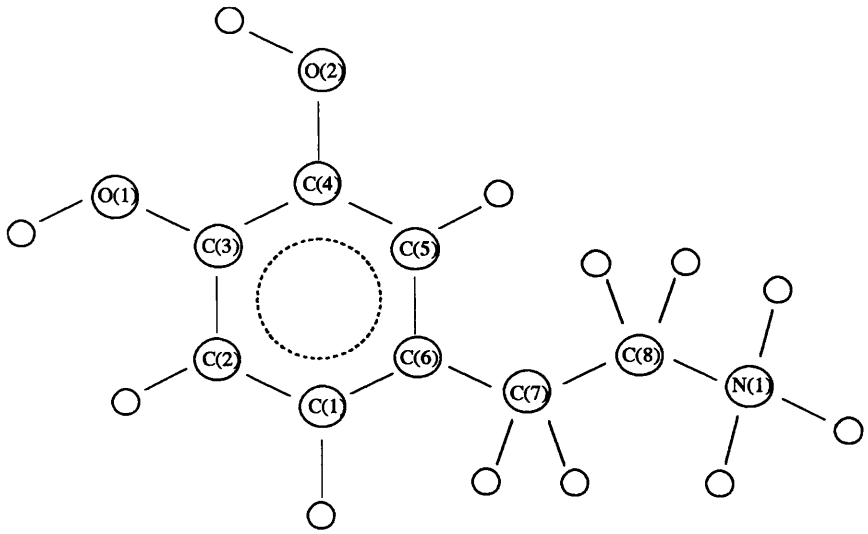


Figure 2.2.1 Schematic representation of the dopamine (DA) agonist

2.2.1 Crystallographic Studies

The Cambridge Structural Database or its popular title: CSD (Allen and Kennard, 1983), contains 114,924 organic and organo-metallic compounds (version 2.3.6 as at March, 1994). Three structures of DA are stored in the CSD. The data set supplied by Bergin and Carlström (1968) did not include hydrogen atom positions as at that time detectors were poor. Since no H atoms were found in the previous dopamine study, Giesecke (1980) used difference synthesis based on Bergin and Carlström's (1968) co-ordinates and seven out of the twelve H atoms were located. After three refinement cycles the remaining five H atoms were located from a new difference map. Giesecke's (1980) co-ordinates are therefore suitable for energy minimisation (equivalent to optimisation of atomic positions) using an all-atom force field without any need to generate hydrogens. In Giesecke's structure, the torsion angles τ_1, C(1)-C(6)-C(7)-C(8), and τ_2, C(6)-C(7)-C(8)-N(1), are -100.4 (3) and 173.2 $(2)^0$ respectively (see figure 2.2.1 for guidance), which means the aliphatic chain is almost fully extended and essentially in the *trans* (α) conformation, forming a plane that is nearly orthogonal to the plane of the ring.

The third and most recent crystal structure determination (Klein,1991) used a least-squares refinement procedure in which multipole parameters were added to describe distortions of

the atomic electron distributions from spherical symmetry. In fact great emphasis was placed on studying the dopamine structure (dopamine hydrochloride) from the viewpoint of charge distribution with the belief that this plays an important role in receptor recognition and binding. The most electron-rich regions of the molecule were the two hydroxyl groups especially at the position of the nonbonded lone pairs. Significantly from the viewpoint of the stated objectives of this Ph.D. study: Klein concluded that any model for the DA receptor must accommodate the distribution of charge and hydrogen bond donors and acceptors in the DA molecule. Klein's crystal structure was also the *trans*-α form of DA.

2.2.2 Which Conformation of Dopamine Is Important At The Receptor Site?

It was thought that the D_1 and D_2 receptors preferred different conformations of the DA molecule (reviewed by Horn, 1990). This seemed a reasonable proposition at the time given that DA can theoretically adopt several distinct conformations: (i) DA can exist in a *trans* or two *gauche* forms (figure 2.2.2); (ii) there are two possible extremes for the *trans* form, i.e. with the catechol ring perpendicular to the $-CH_2CH_2NH_2$ bond (*trans*-α) or coplanar to it (*trans*-β); (iii) when the catechol ring is coplanar to the side chain (*trans*-β), there are two further possibilities depending on the orientation of the ring, i.e. the α and β rotamers.

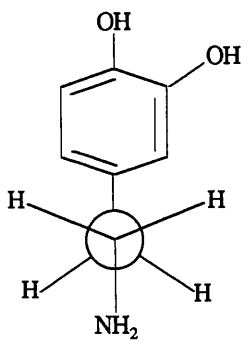
With regard to (i), there are theoretical studies which indicate a preference for *trans* and others which suggest the *gauche* form is preferred. Quantum-chemical calculations and NMR studies by Bustard and Egan (1971) have shown that in aqueous solution at room temperature the *trans* conformation ($\tau_2 = 180^0$) is of lowest energy, but that appreciable amounts of the *gauche* conformation ($\tau_2 = \pm 60^0$) are also present[1]. Pharmacological evidence using rigid and semi-rigid DA analogues shows convincingly that the receptor-preferred conformation is a *trans* species (Horn, 1990). But it is quite ridiculous to believe that DA exists only in the *trans* form. Given that the energy difference between the two forms is only 4-8 kJ mol^{-1} (Horn, 1990; Park, *et al.*, (1992), it seems quite possible from a molecular dynamics viewpoint that the DA ligand might oscillate between the two species.

[1] Various groups have used these findings to design compound sets that match each of the three low-energy conformations of dopamine, for example: Van Drie *et al.*, 1989.

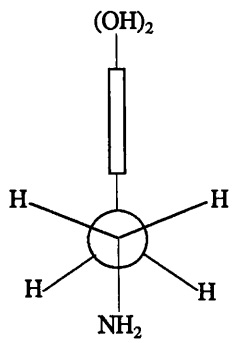
Indeed, Dahl *et al.* (1991a, 1991b) have clearly demonstrated in a molecular dynamics simulation that DA in solution is oscillating on a femtosecond (10^{-15} s) time scale. The aliphatic side chain of the DA molecule moved several times (during a 80-psec molecular dynamics simulation) from one side of the catechol ring plane to the other side and fluctuated rapidly between various *anti* and *gauche* conformations. This is particularly pertinent given that molecular dynamic simulations of transmembrane (TM) helices (e.g. Sankararamakrishnan *et al.*, 1991) with a middle proline indicate that the DA receptor is itself undergoing significant conformational change on a picosecond time scale due to middle prolines located in TM5, TM6 and TM7 in the DA family of receptors; see chapters 8 and 9.

With regard to the possibilities suggested by (ii), Horn points out that it is difficult to answer the question whether the catechol ring is perpendicular (*trans*-α) or coplanar (*trans*-β) to the $-CH_2NH_2$ bond at the receptor site. All three crystal structure determinations of DA are clearly the *trans*-α form. The pharmacologically active analogues 6,7-di-OHATN and apomorphine are both planar molecules and so are related to the *trans*-β instead of the *trans*-α conformer. The potential energy difference between the two species is quite small and so it is quite possible that DA oscillates between both forms and only with regard to rigid analogues of DA is the question of the existence of the *trans*-α. or *trans*-β species significant.

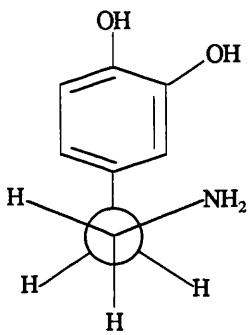
With regard to whether the preferred conformation of the catechol ring is α or β rotamer types, the experimental evidence is also not clear. Catechol (5,6- and 6,7-diOH) semi-rigid derivatives of 2-aminotetralin (ATN; figure 2.2.2) are analogues of *trans*-β form of DA. The β rotamer (i.e. 6,7-diOHATN) derivative is far more potent as a DA receptor agonist than the α rotamer analogue (i.e. 5,6-diOHATN). However, as Horn (1991) also points out: some workers are of the view that the α rotamer analogues are more potent DA receptor agonists using different test methods. Again, it is possible that rapid changes in the conformation of the binding pocket within the hydrophobic core of the dopamine family of GPCRs is spinning confusion amongst the legions of pharmacologists working in this area.



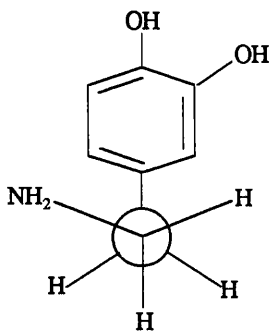
Trans-α



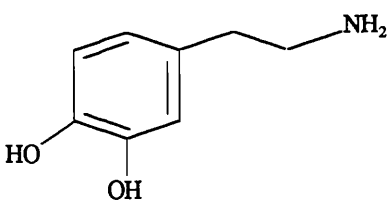
Trans-β



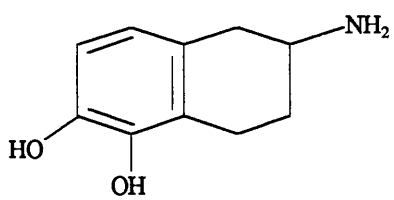
Gauche (+60°)



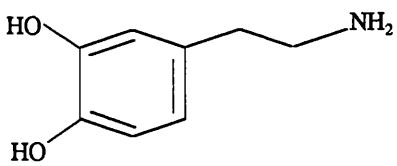
Gauche (-60°)



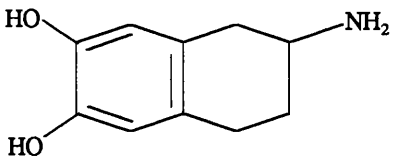
α rotamer



5,6-DiOHATN



β rotamer



6,7-DiOHATN

Figure 2.2.2 Newman projections of the *trans*-α, *trans*-β and *gauche* forms of dopamine; α and β rotamers are two extreme structural forms of the trans-β form of dopamine. 5,6-DiOHATN and 6,7-DiOHATN are analogs of α and β rotamer forms of dopamine.

2.2.3 Electronic Properties of Dopamine

Charge distribution and molecular electrostatic potential fields are of central importance in understanding drug-receptor interactions, particularly if the interactions are largely electrostatic in origin (Waters *et al.*, 1988; Wess *et al.*, 1990). DA has a catechol ring with two hydroxyl groups at *meta* and *para* positions relative to a short aliphatic side chain with a terminal N^+ atom. Hence, three distinct types of electrostatic interactions are likely: hydrogen bonding, ionic bonding and π–δ electrostatic interactions. The binding of ligands to the GPCRs do not involve covalent bonding accept for one notable exception: rhodopsin (Rh), the 11-*cis*-retinal forms a Schiff base with a lysine on helix VII (Smith *et al.*, 1987)[2]. Though the covalent bond can be broken and the retinal released, it spontaneously re-binds to the native protein.

2.2.3.1 Quaternary Positive Charge - potential for cation-anion interaction

The N^+ atom at the terminus of the aliphatic side-chain of DA is liberally quoted as being essential for both agonist and antagonist activity. However, using the Geister-Huckel method (see methods) for calculating charges in molecules it is clear that the N^+ description applied to the quaternary nitrogen is not quite true. The formal positive charge associated with the quaternary nitrogen in the principal valence structure is distributed among the adjacent hydrogen atoms - see figure 2.2.3.1.

Thus the three hydrogen atoms form a larger ball of spread-out positive charge with a smaller positive charge on the nitrogen atom[3]. This phenomenon is also found in acetylcholine. In this molecule, which is the natural agonist of the muscarinic GPCRs, the N^+ atom is nearly neutral. The positive charge associated with the quaternary nitrogen is delocalised over the entire cationic head group, and the nitrogen atom due to its intrinsic

[2] Lysine on helix VII of bRh also forms a Schiff base with a 11-*cis*-retinal molecule - however, this does not constitute significant homology between Rh and bRh.

[3] Experimental evidence that the charge for the ammonium group is not consistent with the conventional formal charge of +1.0 comes from Klein (1991). Klein calculated that the net charge of the ammonium group was +0.2.

electronegativity is in fact the least positive (Beveridge and Radna, 1971). Obvious candidates for providing a "nitrogen-binding site" include the acidic side-chains of aspartate and glutamate.

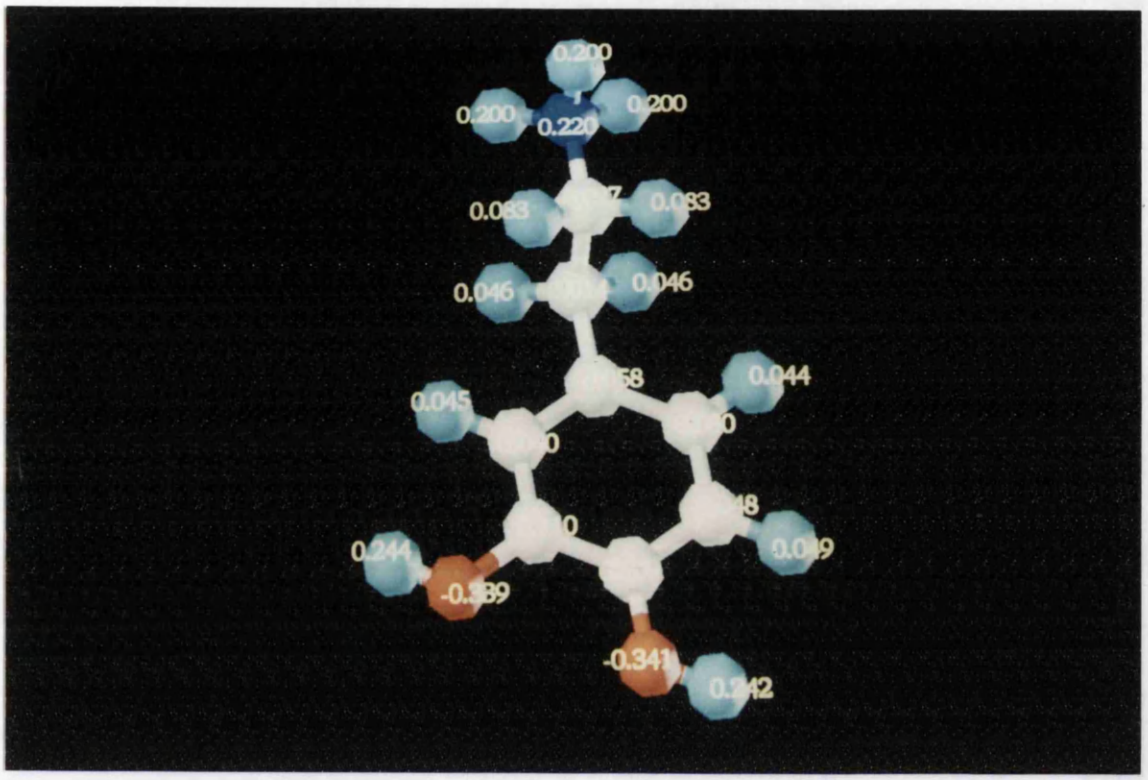


Figure 2.2.3.1 Charge distribution around dopamine calculated using the Geister-Huckel method.

2.2.3.3 π - σ interactions

Seeman's (1980) topographical model for the DA receptors refers to a "hydrophobic site" - which is presumed to be important for hydrophobic interactions with the aromatic ring of the natural agonist. However, this view is rather simplistic. Contrary to the general perception that aromatic interactions are largely hydrophobic in nature, recent studies (for review see Burley and Petsko, 1989) suggest that the aromatic moiety incorporated into the catechol ring are involved in electrostatic interactions with the side-chains lining the binding pocket of DA receptors.

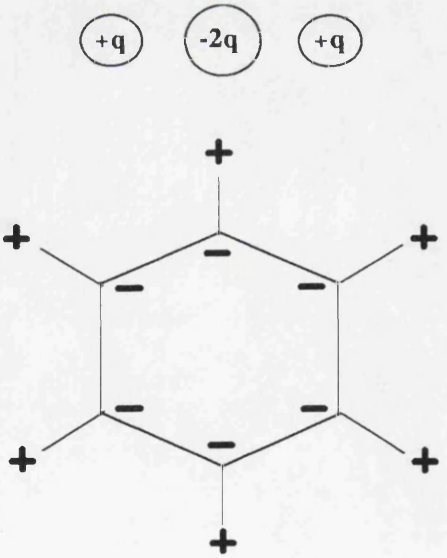


Figure 2.2.3.3 Schematic drawing of a quadruple moments in aromatic ring

Figure 2.2.3.3 clearly shows that aromatic rings have substantial electronic quadruple moments. The simplest aromatic ring can be thought of as consisting of three electronic quadrupoles with common centres and offset from one another by 120°. This leads to several important interactions:

- edge-to-face interactions between aromatic amino acid side chains, which bring a δ(+) hydrogen atom of one aromatic ring near to the δ(-) π-electron cloud of the other aromatic ring;

- oxygen-aromatic interactions (e.g. with Asp, Glu), which bring δ(-) oxygen atom near to the δ(+) hydrogen atoms of an aromatic side chain;

- sulphur-aromatic interactions (e.g. Cys - Phe), which bring the δ(-) sulphur atoms of cysteine and methionine near to the δ(+) hydrogen atoms of an aromatic ring (Reid et al., 1985);

- amino-aromatic interactions (e.g. Lys, Asn, Arg), which bring a positively-charged or δ(+) group near to the δ(-) π-electron cloud of an aromatic moiety. Evidence for amino-aromatic interactions came from a crystallographic study of the interactions of drugs with human haemoglobin which showed the amino group of an Asn residue pointing to the centre of the benzene ring of one of the drugs, suggestive of a hydrogen bond (Perutz et al., 1986; Levitt and Perutz, 1988).

- aromatic-quaternary ammonium ion interactions. For example, the 3D structure of acetylcholinesterase form *Torpedo californica* electric organ has been determined by x-ray analysis to 2.8 Å resolution (Sussman *et al.*, 1991). Subsequent modelling of acetylcholine binding to the enzyme suggested that the quaternary ammonium ion is bound to some of the 14 aromatic residues which line a gorge. A preferential interaction between

the positively charged quaternary nitrogen with the π-electrons of the aromatic side-chains clearly exists. Similar observations were made by Dougherty and Stauffer (1990) who noted that the neurotransmitter acetylcholine binds to a completely synthetic receptor (host) comprising primarily of aromatic rings. Again, the stabilising interaction clearly appeared to be between the quaternary ammonium ion and the electron-rich π systems of the aromatic rings (cation-π interactions).

These interactions are both ubiquitous and numerous and together with hydrogen bonding and charge-charge interaction, play a major role in molecular recognition mechanisms with regard to ligand-receptor binding (Verdonk et al., 1993). In particular, they are likely to play an important role in the binding of selective agonists and antagonists to the DA receptors. For example, a recent study of the contribution of the 1-phenyl substituent to the molecular electrostatic potentials of some benzazepines in relation to selective DA D_1 receptor activity (Pettersson *et al.*, 1992) clearly demonstrated that an important part of the interaction between the phenyl ring in the benzazepines and the receptor is due to electrostatic forces.

2.2.3.4 Potential For Hydrogen bonding

It is clear that the two hydroxyl groups which characterise the catechol moiety of the natural ligand are available for donor/acceptor hydrogen bonding and at least one hydroxyl group is required for successful agonist induced conformational change in the catechol amine binding GPCR[4]. Indeed, if both of the hydroxyl groups are removed from the aromatic ring - this converts the natural agonist into a potent antagonist (Civelli *et al.*, 1992). Hence, hydrogen bonding is essential for agonist activity. The importance of

[4] Extensive structure-activity relationship (SAR; Cannon, 1985) studies have shown that only the *meta*-hydroxyl group is essential for "bioactivity" (i.e. binding). Cannon's findings have prompted Van Drie *et al.* (1989) to design D_2 agonist phenols rather than catechols. It has transpired that the *meta*-hydroxyl interacts with Ser-204 while the *para*-hydroxyl hydrogen bonds with Ser-207 (re: β_2 receptor (a catechol amine binding GPCR; Strader *et al.*, 1989a,b). However, Strader concluded that either hydroxyl groups of the phenyl ring hydroxyls ensured efficacy. [β-receptors (β_1, β_2 and β_3) have significant homology with DA receptors.]

hydrogen bonding is amply demonstrated by the fact that all the topographical models of the DA receptors include sites for hydrogen bonding; for example, Seeman's (1980) model - see figure 2.2.3.4. Residues such as Ser, Thr, Asn or Gln are obvious candidates as "sites" for hydrogen bonding in the binding pocket of DA receptors; the polar oxygen atoms of tyrosine and even accessible carbonyl backbone oxygen atoms may also provide opportunities for hydrogen bonding with DA. That hydrogen bonding does play an important part in the binding of the agonists to DA receptors is not surprising given the nature of the hydrogen bond. Because of its small bond energy and small activation energy involved in its formation and rupture, the hydrogen bond is especially suited to play a part in reversible receptor binding mechanisms.

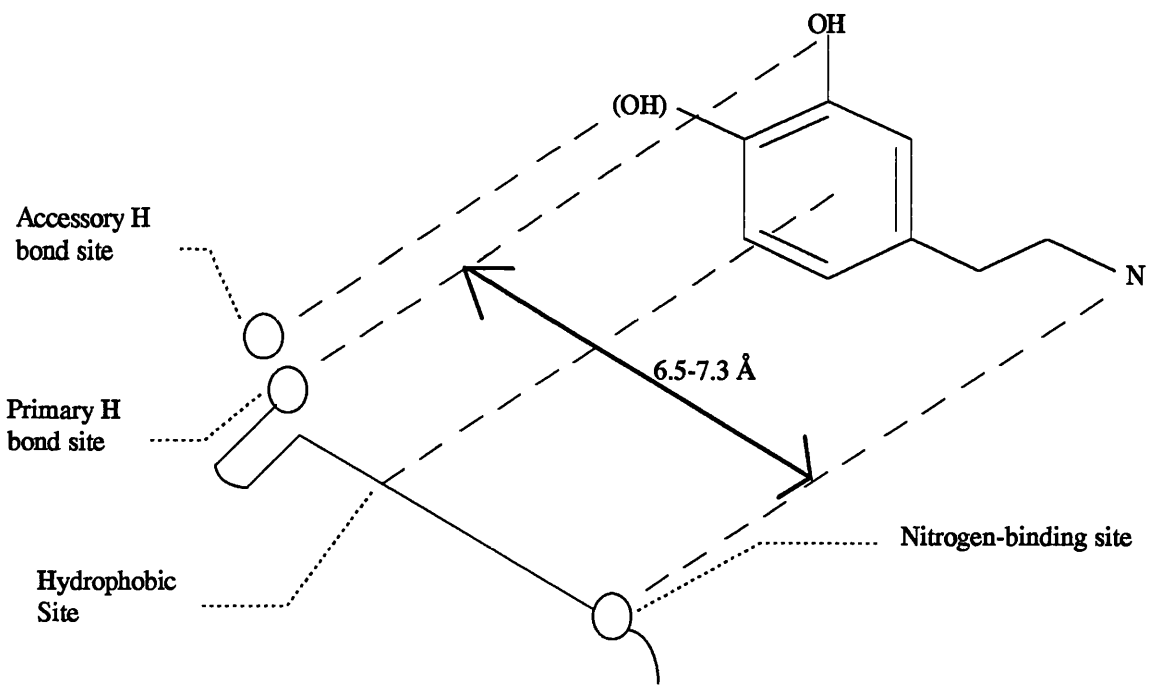


Figure 2.2.3.4 Model for DA receptors - adapted from Seeman's (1980) topographical model for the DA receptors. It is clear that hydrogen bonding plays a big part in the agonist activity of the natural ligand dopamine.

3. The Dopamine G Protein Coupled Receptors (GPCRs)

3.1 Role Of The Dopamine Receptors In Signal Transduction

The dopamine (DA) receptor (like all other GPCRs) forms part of a critical message transduction pathway across the cell membrane of the host cell. In the classic agonist-receptor interaction model (reviewed in Neal, 1989; see figure 3.1) the agonist binds to the receptor and the agonist/receptor complex can bind to the transducer molecule to form the agonist/transducer complex to initiate the desired response. If the DA GPCR follows the classic agonist-receptor interaction model then the agonist should bind to the DA GPCR and in response to agonist binding conformational changes would occur in the receptor molecule to allow a second molecule (an appropriate G protein) to bind to the intracellular side of the receptor protein.

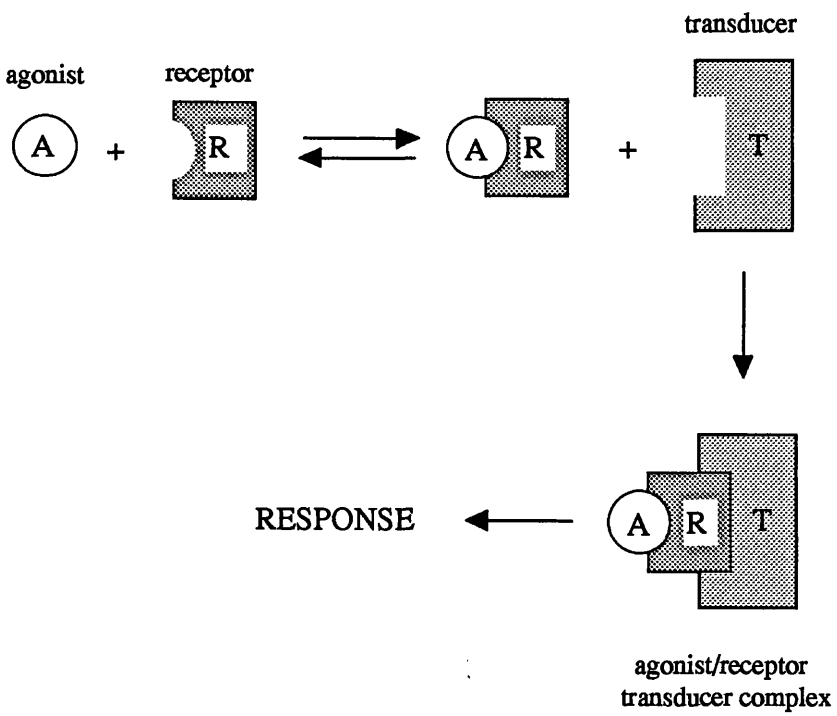


Figure 3.1 Classic model of agonist-receptor interactions
adapted from Neal, 1989.

However, the findings of Urwyler and Markstein (1986) showed experimentally that the DA receptors do not follow the classic drug-receptor interaction model. They showed, prior to the actual cloning of D_3 and D_4, that "D_3" and "D_4" binding sites are high agonist affinity states of the D_1 and D_2 receptors respectively. These workers even went so far as to suggest that: "It may well be that the multiplicity of dopaminergic recognition sites reflects simply the fact that dopamine receptors can exist in different states, with different affinities especially for agonists." Earlier other workers had suggested that the "D_3" binding site was actually a high agonist affinity state of the D_1 receptor (e.g. Seeman, *et al.*, 1985).

This deviation from the classic drug-receptor model is better explained in terms of the "ternary complex model" (figure 3.1.1) which was originally proposed for the β-adrenergic receptor (De Lean *et al.*, 1980). This model was applied to the dopamine receptors (Sibley *et al.*, 1982; Wreggett and De Lean, 1984). In this model, the receptor (R in figure 3.1.1) can not only interact with the ligand (an agonist or an antagonist), but also with an additional membrane component, the guanyl nucleotide binding protein G. The coupled form RG has a high affinity for agonists, whereas the free receptor R, would correspond to the low agonist affinity state. Agonists tend to stabilise the interaction between R and G in the form of the ternary complex LRG; however, in this form the G-protein has a high affinity for GTP and on binding of this nucleotide the complex dissociates and agonist affinity is reduced.

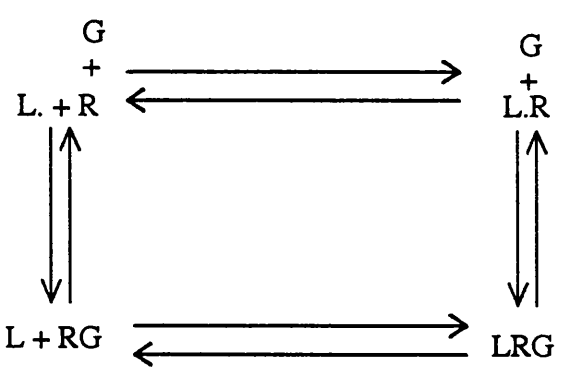


Figure 3.1.1 The ternary complex model for drug-receptor-G protein interaction (adapted from Wreggett and De Lean, 1984).
Abbreviations: R, receptor; L, ligand; G, G-protein.

D_1 receptors couple to G_s proteins which stimulate adenylate cyclase activity. In contrast, D_2 receptors couple to G_o (which have no effect on or inhibit adenylate cyclase activity) or G_i proteins (which inhibit adenylate cyclase activity). However, this description of events is a simplification - Niznik and Van-Tol (1992) instead explain the role of the five distinct genes recently cloned in terms of their use as "tools for molecular psychiatry". While the role of the dopamine receptors in some psychoses (and Parkinson's disease) is generally accepted, though not well understood, there is a growing body of experimental evidence which suggests the dopamine receptors play a far more diverse and interactive roles in the functioning of higher organisms.

In their review of the functional implications of the D_1/D_2 classification scheme Clark and White (1987) looked at the pharmacological effects of compounds exhibiting putative selective agonist or antagonist activity at DA receptor sites. Their study supported the view that D_1 and D_2 receptors interact in both an opposing and synergistic fashion in the control of behavioural expression. Experimental work by Benkirane *et al.* (1987) showed that in rat *substantia nigra,* exogenous and endogenous dopamine causes inhibition of serotonin release by activation of D_1 subtype receptors. The inhibition of serotonin release was antagonised by SCH 23390 (a D_1 selective antagonist), but was unaffected by S-sulpiride (a D_2 selective antagonist).

Jose *et al.* (1992) have identified a D2-like receptor, which they have named: D2Ak. It is a renal dopamine receptor with some pharmacological features of the D2 receptor but not linked to adenylate cyclase. Found in the inner medulla, D2Ak is linked to the stimulation of prostaglandin E2 production, apparently due to stimulation of phospholipase A2. The DA1 receptor in the kidney is associated with renal vasodilatation and an increase in electrolyte excretion. The role of the DA2 receptors in the kidney requires clarification.

Of particular concern, are the recent experimental findings of Johnston *et al.* (1993). These workers noted that abnormalities in D_2 receptors my be implicated in the development of some pituitary tumours. This finding ties up nicely with the completely independent work of Wong *et al.* (1993a; also: 1992, 1993b,c) who have in *vitro* and in *vivo* evidence that

DA, by acting through D_1 receptors in the pituitary, functions as a growth-hormone releasing factor in goldfish. Finally, somewhat mirroring the work of Clark and White (1987), Wong *et al.* (1991) has reported that the dopamine receptors D1 and D2 (also GABA(A) receptors) play a major role in modulating the control of motor function by the nucleus-accumbens of rat brains.

Elucidating in more detail the manner in which the functioning of the dopamine receptors interact with other potent agonists and their receptors and parts of the endocrine system will hopefully lead to potent drug treatments which lack side-effects. This confusion simply arises due to our lack of understanding of the interactive role played by the dopamine receptors, both with other dopamine receptor subtypes, receptors of other GPCR families and even parts of the endocrine system.

3.2 Overall Topology

Integral membrane proteins contain one or more membrane-spanning segments. Excluding the porins (membrane channel proteins, for example see Weis and Schulz, 1993) which contain intramembrous β-sheet as their major structural component, TM segments are stretches of hydrophobic amino acids of sufficient length to span the lipid bilayer as α-helices. Bacteriorhodospin from *Halobacterium halobium* is composed of seven such stretches (though the seventh is amphiphilic) aligned to a hepta-helical motif of anti-parallel helices essentially normal to the lipid bilayer (Henderson and Unwin, 1975; Henderson *et al.*, 1993). Given that the hydropathy plots such as the Kyte Doolittle plot (1982) of GPCRs show clear similarities to the hydropathy plot of bRh (see chapter 5) it is generally concluded that GPCRs include a similar heptahelical motif (see figure 3.2)[1]. Indeed, the

[1] The D_2 family (i.e. D_2, D_3 and D_4) differs from the D_1 family (i.e. D_1 and D_5) in two clear ways: 1) members of the D_2 family have a long intracellular loop (iIII) between TM helices V and VI whereas members of the D_1 family have a shorter version of iIII; 2) D_2 sub-type receptors have a short carboxy tail whereas D_1 sub-type receptors have a long carboxy tail. In both families, hydrophobic TM helices are connected by short stretches of hydrophilic amino-acids.

identification, by protein sequencing, of a large number of sites on rhodopsin (a GPCR) which were subjected to chemical or biochemical modification or to proteolytic cleavage, confirmed that rhodopsin traversed the lipid bilayer seven times (reviewed: Findlay and Pappin, 1986; Findlay *et al.*, 1988). Fujiwara *et al.* (1991) have pointed out in their paper ("The Molecular Biology of Dopamine Receptors"); the English translation reads: "Analysis of their (the dopamine receptors) amino acid sequences has shown that these subtypes belong to the G protein-coupled receptor family, with seven TM domains, a C-terminus within the cell and a N-terminus outside the cell." This seems a reasonable statement given that rhodopsin has significant homology with each member of the dopamine family of GPCRs (chapter 5).

However, the question remained: *are the helices arranged in a anti-parallel fashion.* Indeed, Findlay has asked this question (Ryba *et al.*, 1992) and noted that 7! permutations are possible (i.e. 5,040 possible arrangements). However, precisely the same question was considered over a decade ago by Engleman *et al.* (1980) who looked at the path of the polypeptide in bRh. They likewise concluded that 7! permutations are possible. They used a simple algorithm to rule out physically implausible arrangements (e.g. by looking at lengths of hydrophilic stretches of amino-acids between the then putative TM α-helices) to produce only one physically plausible arrangement: the anti-parallel heptahelical motif. This basic motif was confirmed by Henderson *et al.*, 1990 by modelling the structure of bRh based on high resolution electron cryo-microscopy - a technique developed by Henderson. It is generally agreed that the TM component of GPCRs consist of seven TM alpha helices arranged in an antiparallel fashion essentially like that demonstrated for bRh. The question of the exact location of the GPCR helices and their relative tilts still remains to be resolved.

Gebhard *et al.* (1993) provided a projection map of Rh showing the configuration of the TM helices. A projection of Rh (9Å resolution) and for comparison a projection of bRh at 7Å resolution was also obtained using the same method (see figure 3.2.1). The projection density map for Rh shows a drawn out arc-shaped feature and four resolved peaks of density. They interpreted the four peaks to be four TM helices orientated nearly perpendicular to the lipid bilayer. The arc is thought to represent three tilted helices

(making seven TM helices in all). Significantly, they noted that: "the projection structure of rhodopsin is less elongated and slightly wider and the helices are tilted differently from bRh". In her paper on the probable arrangement of the helices in GPCRs, Joyce Baldwin (1993) also noted that: "the structure of rhodopsin seen in this map is clearly different from the structure of bacteriorhodospin". However, a full 3D structure is still required. Only with a 3D structure can details be accurately evaluated concerning tilts and orientations of the helices relative to one another, the retinal moiety and the surrounding lipid. It is generally agreed that helices are orientated so that conserved and highly conserved residues face inwards and away from the lipid bilayer[2] - this topic is considered in detail in chapter 8. Also that charged residues such as aspartate or glutamate would never be exposed to the surrounding lipid bilayer. However, Maloney-Huss and Lybrand (1992) found that a glutamic acid side chain (E122) in helix III in their model of β_2 adrenergic receptor protein is exposed to the surrounding lipid. Likewise, Fourier analysis (chapter 8) suggests that some conserved residues do not always face inwards or towards other helices. This suggests that helix shear re-orientates specific helices to allow these residues to play their role during the binding and release of the agonist ligand - more details in section 3.4.3 and chapter 8.

[2] This might be expected from biostructure analysis work carried out by Bordo and Argos (1990) who concluded that a residue buried in a protein core would have different mutational constraints than one lying at or near the protein surface (in the case of the TM components of integral membrane proteins the *protein surface* are the sides of the helices facing the lipid).

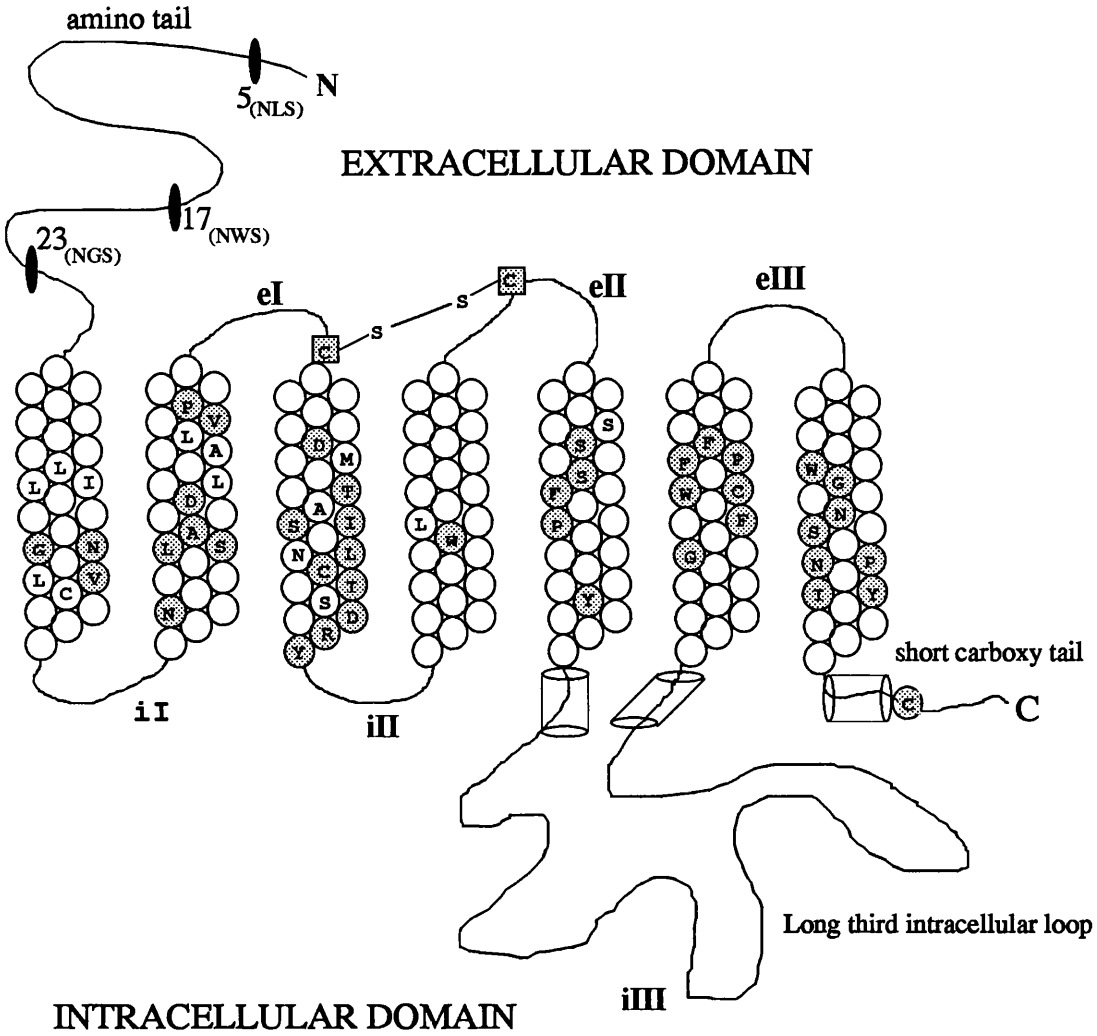


Figure 3.2; Schematic drawing of the overall topology of the dopamine (DA) receptor. Single letter amino acid code is used. NLS, NWS and NGS are tripeptide glycosylation sequences in D_2; 5, 17 and 23 are the positions of N in each of the tripeptide sequences. Residues in clear circles are conserved throughout the dopamine family. Those in shaded circles are conserved throughout a representative set of cationic amine G protein coupled receptors: dopamine (D_1 through to D_5), β-adrenergic (β_1, β_2 and β_3), α-adrenergic (α_{2A}, α_{2B} and α_{2C}) and the serotonin $5HT_{1A}$ receptor (see appendix 1). The residue assignments follow Baldwin's scheme, e.g. there is an aspartate residue in transmembrane helix 2 (TM2) at position 14 (Asp_{214}) and in TM3 at position 7 (Asp_{307}) - see figure 3.4.1. The cylinder outlines at the amino and carboxy termini of iIII signify short intracellular α-helices - these have been predicted using the secondary prediction algorithm PHD (Rost and Sander, 1993 -see chapter 5) and are in agreement with the findings of Maloney-Huss and Lybrand (1992). A α-helix (3 turns) is also predicted using PHD between carboxy-end of TM7 and a conserved Cys residue as shown.

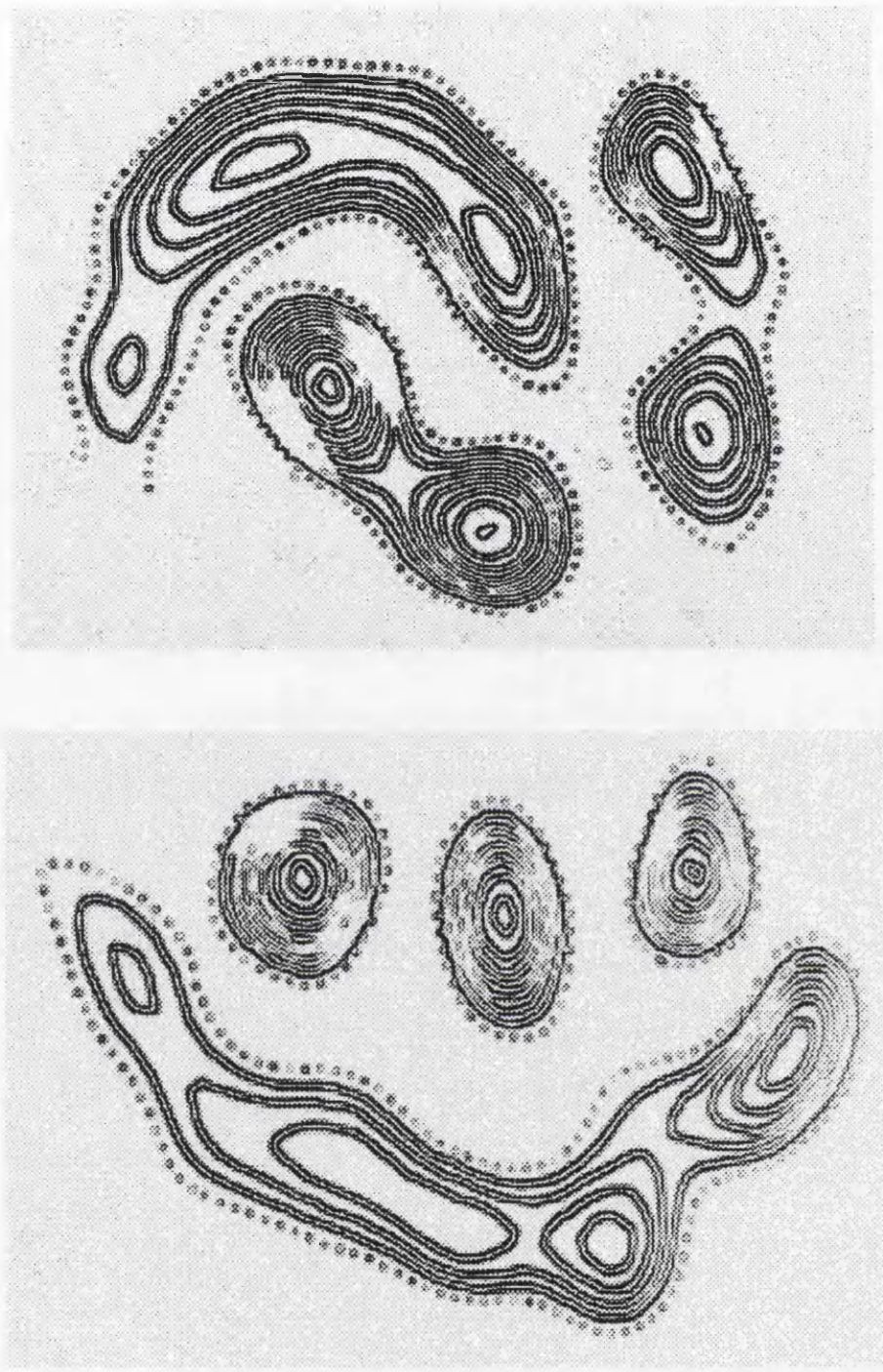


Figure 3.2.1; Projection density of a single molecule of rhodopsin at 9 Å resolution (top) and bRh at 7 Å resolution (bottom) viewed from the intracellular side of the membrane - Gebhard *et al.* (1993).

3.3 Tripeptide Glycosylation Sequences in GPCRs

It had been noted some years ago that in N-linked glycoproteins the Asn bearing the carbohydrate chain was part of the tripeptide sequence -Asn-X-Ser or -Asn-X-Thr (i.e. -Asn-X-Ser/Thr; NXS/T using single letter code), where X represents one of the 20 amino acids (Eylar, 1965; Marshall, 1974). However, two caveats soon emerged: 1) not all proteins containing this sequence are glycosylated; 2) X can be any of the 20 amino acid except perhaps Asp (Marshall, 1974). With regard to the first caveat, the factor determining whether a protein with the -Asn-X-Ser/Thr is glycosylated depends on where this tripeptide sequence is situated. If the sequence is accessible to the oligosaccharide transferase (i.e. on exposed coils or β-turns) the protein is glycosylated.

In addition, the polar hydrophilic sugar moieties of these proteins (termed glycoproteins) invariably are found exposed to the exterior medium and are thus available for interaction with external factors (Cotmore *et al.,* 1977). In fact, thermodynamic considerations had suggested that the hydrophilic carbohydrates should be found in the hydrophilic medium (Singer, 1971; Singer and Nicolson, 1972) rather than the hydrophobic bilayer. However, the fact that glycosylated sites are found in the exterior medium was demonstrated in the case of the red cell major glycoprotein (PAS-1, glycophorin) where the sialic acid containing moieties were experimentally shown to be located in the N terminus of the protein external to the cell (Tomita and Marchesi, 1975; Cotmore, 1977). The C terminus being located on the cytoplasmic side of the cell. Indeed, Nicolson and Singer (1974) noted that for a number of systems the *carbohydrate chain* is located on the external side of the membrane. Every member of the DA family of GPCRs has potential sites available for glycosylation (see table 3.3 below). For example, the D_2 human receptor has three potential sites for glycosylation (at positions 5, 17 and 23). Hence, the DA receptor family belong to the general class of glycoproteins.

Receptor	Position (of Asn residue)	Tripeptide Sequence (Single letter code)	Location	Extracellular?
D1	5	**NTS**	AMINO TAIL	YES
D2	5	**NLS**	AMINO TAIL	YES
D2	17	**NWS**	AMINO TAIL	YES
D2	23	**NGS**	AMINO TAIL	YES
D3	12	**NST**	AMINO TAIL	YES
D3	19	**NST**	AMINO TAIL	YES
D3	173	**NTT**	eII	YES
D4	3	**NRS**	AMINO TAIL	YES
D5	7	**NGT**	AMINO TAIL	YES

Table 3.3 Position of possible glycosylation sites in the dopamine (DA) receptors.

3.3.1 Genetic Analysis

Strader and Dixon (1992) pointed out that genetic analysis studies of the β-adrenergic receptor (and therefore by implication any GPCR) are directly applicable to the entire family of GPCRs - including the DA family of receptors. In their review of genetic analysis of the β-adrenergic receptor they identified three approaches: site-directed mutagenesis, deletion mutagenesis and construction of chimeric proteins. A fourth approach is implicitly suggested by Patrick Argos (Bordo and Argos, 1990) who noted that point mutations of structural equivalent residues in families of known structures are equivalent to *in vivo* site-directed mutagenesis. A similar study of natural protein engineering and design of the DA receptor family must wait until a collection of high resolution 3D structures are available.

However, using cloning technology it is possible to study allelic polymorphism in any DA receptor. Van Tol *et al.* (1992) did just that in their study of D_4 variants in the human population and suggested that the occurrence of variable repeat regions in the third

intracellular loop[3][4] (iIII) might account for observed phenotypic variability in psychotic patients on clozapine treatment. In a more limited study (Nishimot and Okamoto, 1992; Ikezu *et al.*, 1992) analysis of primary structure of the muscarinic and α_2-adrenergic receptors noted that only certain regions of iII and iIII are responsible for binding to the G-proteins. These regions did not correspond to the general location of any of the repeat regions observed by Van Tol's group. In addition, iIII of D_4 (or any of the DA receptors) is not directly involved in the binding of the antagonist clozapine since it is an intracellular loop. Also, numerous workers (e.g. Dixon *et al.*, 1987a and 1987b) have noted that residues directly involved in binding of ligands are found in the hydrophobic (i.e. TM component) part of GPCRs. Shaikh *et al.* (1993) also looked at the frequency of occurrence of dopamine D_4 receptor subtypes (D4.2, D4.3, D4.4, D4.5, D4.6 and D4.7) in clozapine-treated patients and their response to clozapine. They found no significant differences in allele frequencies between responders and non-responders[5]. It is therefore very doubtful that variation in simple repeat homology in the middle of iIII of D_4 could have such dramatic effects.

Genetic analysis as envisaged by Strader and Dixon (1992) cannot be contemplated unless the genes of interest have been cloned. Unfortunately classical methods of amino acid sequencing are generally difficult for membrane proteins, particularly for the hydrophobic regions. Classical methods require the purification to homogeneity of the gene protein product, which has not been achieved for e.g. the D_1 receptor (Niznik *et al.*, 1992). However, use of the polymerase chain reaction (PCR) has proved very successful, to the extent that Watson and Arkinstall (1994) were able to compile a book containing nearly 300 pages listing GPCR primary structures.

Cloning gives the first opportunity to characterise the receptors. For example, Selbie *et al.* (1989) performed a molecular analysis of the human D_2 receptor and Sokoloff *et al.* (1990)

---

[3] Van Tol's group discovered that a 48-base-pair sequence in the third intracellular loop of D_4 exists either as a direct-repeat (D4.2), as a fourfold repeat (D4.4) or as a sevenfold repeat (D4.7).

[4] The author had previously noted the existence of a direct-repeat (i.e. D4.2) using a simple algorithm designed to detect internal homology within GPCRs - see chapter 6.

[5] They did add the caveat that the relationship between particular alleles or genotypes and more subtle differences in clozapine response require a larger sample of patients to perform the necessary statistical evaluation.

used molecular-cloning as a vehicle to characterise the D_3 receptor. Applying hydropathy analysis to the primary structures allowed these workers to make one important distinction which distinguishes the D_1 family from the D_2 family. D_1 receptors have a long intracellular carboxy tail and a short loop iIII whereas iIII in D_2 receptors is much longer and the carboxy tail much shorter. In addition, once the first DA receptor had been cloned (D_2 - Selbie *et al.*, 1989) the subsequent cloning of the other members of the DA receptor family helped focus attention on what were perceived to be key residues. In particular, newly cloned DA receptors were checked to see if they had equivalents to the two Ser residues (Ser-194 and Ser-197) of D_2 which were believed to be involved in hydrogen bonding to the natural agonist DA. Once several cationic amine binding GPCRs had been cloned and consequently sequenced, this allowed a multiple sequence format file to be compiled and conserved Ser residues identified along with other residues suitable as candidates for site-directed mutagenesis experiments.

In analysing the results of mutagenesis experiments, it is vital to be certain to differentiate between functional mutations and those that effect structural integrity of the protein. Otherwise residues which actually play an important structural role may appear to play an important role in binding when in fact the binding would tend to be disrupted anyway as a result of miss-folding of the receptor protein. It is also important to locate those regions of the receptor protein responsible for certain functions which can be determined by using chimeric receptors. Of particular interest are those regions of the DA receptor which play a decisive role in binding agonists and antagonists. Unfortunately, because of the enormous commercial interest in the DA receptors, there are very few papers on the genetic analysis of binding of DA GPCR agonists and antagonists. This mirrors the problem with regard to protein modelling of the DA receptor, where 3D models (particularly their x, y, z atom co-ordinates) do not find their way into the public domain (Humbler & Mizadegan, 1992).

However, with regard to the natural agonist (DA) as discussed in earlier sections, the binding pocket of the DA receptor should contain a counter-ion to the cationic amine group (present in all DA agonists and antagonists), side-chains capable of hydrogen bonding to the hydroxyl groups of the catochol (particularly the *meta* hydroxyl group) plus side-chain capable of π-σ or π-π interactions with the aromatic ring of DA, i.e. Met,

Cys, Arg and Asn (π-σ interactions); Tyr, Phe and Trp (π-π interactions). Antagonists differ from agonists in one fundamental respect: they lack the ability to interact with the hydrogen bonding sites in the binding pocket. Of enormous interest are chimeric studies involving the DA receptors. Of particular value would be the construction of chimeric dopamine receptors, in which various regions from two dopamine receptor subtypes are combined and the pharmacological phenotype of the resulting hybrid protein analysed. For example, the experiments performed by Frielle *et al.* (1988) who constructed chimeric β_1/β_2-adrenergic receptors to determine the structural basis of β-adrenergic receptor subtype specificity. Such experiments could be repeated on the DA receptors to locate regions of the hydrophobic core of each receptor that distinguish it from other DA receptor subtypes. Such information would be of enormous value as it would aid medicinal chemists in developing drugs which bind to e.g. D_1 in preference to D_2. But as explained above, this information is not in the public domain.

3.4 The Binding Mechanism of the Dopamine Receptor

The likely role of various side residues and non-covalent forces in binding agonists (and antagonists) has already been outlined. Here three important aspects which are likely to play a decisive role in binding agonist ligands is described: the location of the binding pocket and binding equilibria, α-helix kinking and helix interface shear mechanism. Of course, these important aspects are taken up in the results and discussion section where they are discussed in detail in the context of the 3D models of the DA receptors created by the author.

3.4.1 Location Of Binding Pocket And Binding Equilibria

Saunders and Findlay (1990) carefully analysed the findings of Strader *et al.* (1989b) who carried out a detailed genetic analysis of the structure function relationships of catechol amine binding GPCRs. Saunders and Findlay were particularly interested in the finding that while Asp-113 (β_2 adrenergic receptor) on helix III (10Å from the surface) is

vital for agonist and antagonist binding, the Asn-113 mutant is still fully coupled to cyclase at very high concentrations of agonist. This suggested the possibility that a second acidic residue may function to bind agonist ligands. Asp-79 on helix II some 15Å from the receptor surface is therefore likely to also be involved in agonist as opposed to antagonist binding. They extrapolated these findings to propose a sequence of equilibria that underscore muscarinic receptor activation following agonist binding. Essentially, the agonist first binds to Asp(105) on TM helix III of the muscarinic receptor M_1 and then moves into a position which enables it to interact with the deeper Asp(71) on TM helix II and then mainly with Asp(71). This proposal has been extrapolated to describe a similar set of putative equilibria for agonist binding to the D_2 receptor - see figure 3.4.1.

Dahl *et al.*, (1991b) constructed a 3D model of D_2 from its amino acid sequence and used it to simulate the MD of DA-receptor interactions using a Cray X/MP-28 supercomputer. Their model consisted of seven α-helical TM segments that formed a central core with a putative ligand-binding site. Space between helices II and VI was occupied by low energy conformations of the ligand i.e. *gauche* ($\pm 60^0$). Also, the MD simulation clearly demonstrated that the protonated amino group of DA became orientated toward negatively charged aspartate residues in helix II and helix III in much the same manner as proposed by Saunders and Findlay (1990). In fact, Dahl (Dahl *et al.*, 1991a,b) favours the so called "zipper" mechanism (Burgen *et al.*, 1975) whereby the ligand changes in conformation during the binding process which takes place in several successive steps.

On the basis of earlier work (Strader *et al.*, 1989c) Strader and Dixon (1992) suggested that there are overlapping binding sites for agonists and antagonists. It seems reasonable therefore to also suggest that different agonists also have overlapping binding sites. Sylte *et al.* (1993) have performed MD studies on a model of the serotonin ($5HT_{1a}$) receptor (which has high homology with the DA receptors) and ligands. The ligands used in the MD study were: serotonin (5HT - the natural agonist), buspirone (partial agonist), S(+)-methiothepin and S(-)-methiothepin (both isomers are antagonists), ipsapirone (IPS - high affinity for $5\text{-}HT_{1a}$ receptor). By calculating interaction energies between ligands and specific residues, Sylte showed that up to 22 different amino acid residues may form a ligand binding pocket, and contribute to the specificity of ligand recognition and binding.

For example, the binding pocket clearly differs between 5HT and IPS. The use of supercomputers and appropriate software to characterise binding domains offers spectacular insight and will be of great use to medicinal chemists who seek to design ligands with specific binding profiles.

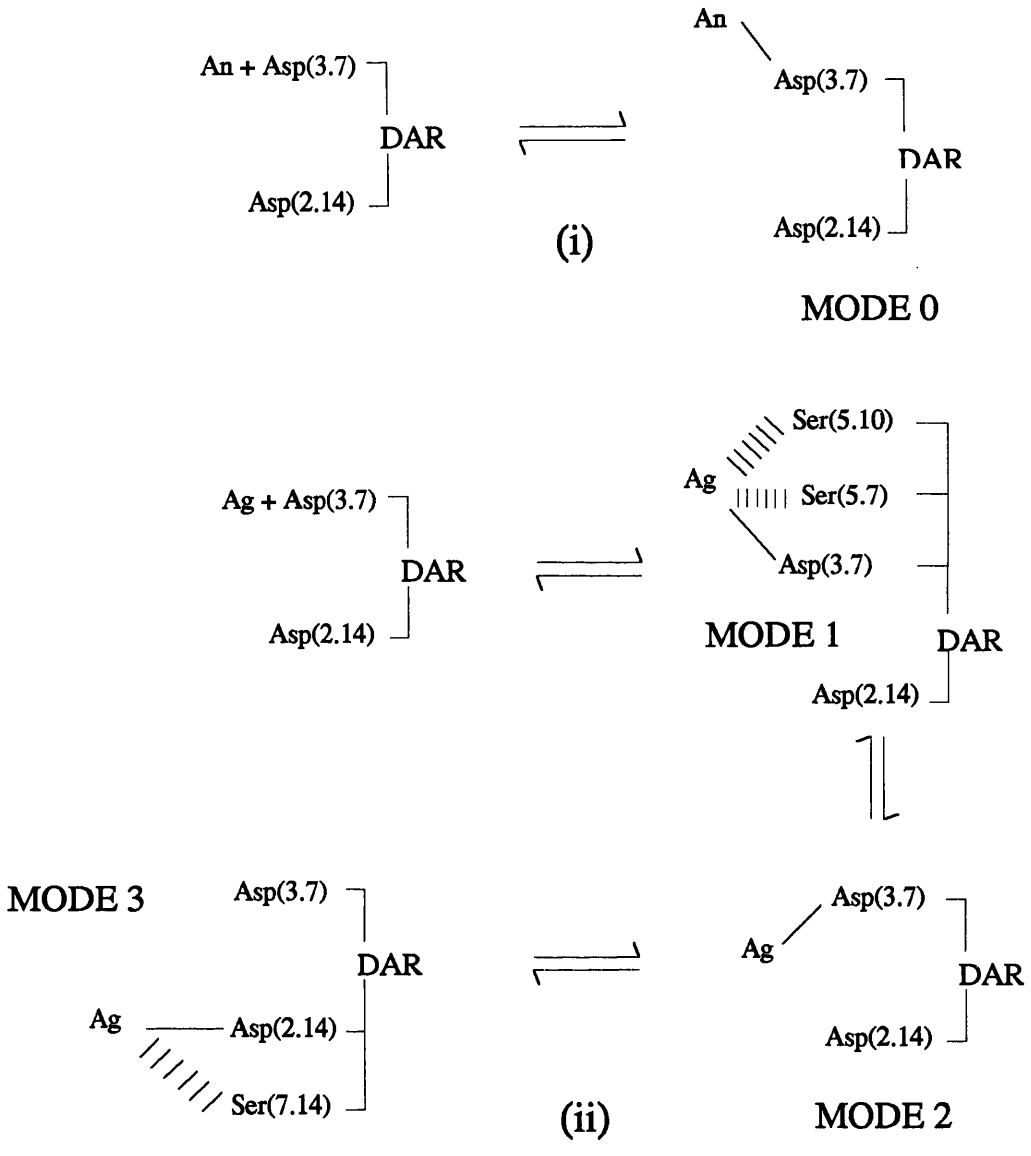


Figure 3.4.1 Simple model of antagonist (An) and agonist (Ag) binding to dopamine receptor (DAR) subtype D_2. (i) antagonist binds to the dopamine receptor. The principle non-covalent interaction being a reinforced ionic bond between the positively charged group on the cationic ligand and the negatively charged Asp residue on helix 3, position 7. This is termed MODE 0. (ii) agonist binding to the dopamine receptor (D_2); hydrogen bonding between the agonist and side chains of Ser residues on helix 5 at positions 7 and 10 together with ionic bonding with Asp on helix 3 - this is termed MODE 1. MODE 1 differs from MODE 0 in one important aspect: there is no hydrogen bonding involving the Ser residues on helix 5 in MODE 0. The agonist moves deeper into the receptor and at some point (MODE 2) lies close to Asp residue on helix 2. Electrostatic interaction between Asp residue on helix 2 characterises MODE 3. However, hydrogen bonding is likely between the catechol moiety and conserved Ser residues e.g. Ser(7.14) which is conserved in the dopamine family of GPCRs. The climax of MODE 3 is the exposure of the DRY (single letter code) motif on C-flank of helix 3 which is rendered available for coupling to G-protein in the intracellular medium thus completing the message transduction process. Adapted from Saunders and Findlay (1990).

3.4.2 α-Helix Kinking

Kinking of TM α-helices caused by middle prolines has been noted by Henderson *et al.* (1990). However, the relevance of detailed conformational structure and molecular dynamics studies of TM helices with a middle proline (Sankararamakrishnan and Vishveshwara, 1990; Sankararamakrishnan *et al.*, 1991) to the binding of agonists by catechol amine binding GPCRs has not been fully appreciated.

For example, Saunders and Findlay (1990) mention the importance of the two conserved serine residues on TM helix V. The fact that helix V also has a middle proline was not considered. Presumably the positively charged quartinary group of the agonist (and antagonists) is electrostatically attracted to the acidic side chain of Asp-113 (β_2 adrenergic receptor). This seems a reasonable conjecture, but it begs the question why only the agonist can approach and then interact with the deeper Asp-79 residue. Only the agonist interacts with the two conserved hydrogen bonding sites (e.g. Ser-204 and Ser-207 in β_2) on helix V. However, this fact by itself does not adequately explain why only the agonist is able to interact with the deeper Asp-79 residue[6]. Interestingly, Dahl (Dahl *et al.*, 1991b) observed during the MD simulation of the docking of the fluctuating DA molecule that "the electrostatic forces were not sufficient to attract the dopamine molecule to the postulated binding site during the simulation" and introduced a slight distance constraint. Hence, something else must be driving the agonist deeper into the binding pocket.

A clear explanation comes from detailed molecular dynamic studies of TM helices with a Pro residue in the middle position. A TM helix with a middle proline oscillated every 2-4 ps between a largely straight structure and a highly bent structure (Sankararamakrishnan *et al.*, 1991). The omega (ρ -1) torsion angle was also found to vary in sympathy with the conformation of the proline residue which fluctuated between a *puckered-up* (where χ_1 is negative) and *puckered-down* (where χ_1 is positive) conformation - see figure 3.4.2 for

[6] MaloneyHuss and Lybrand (1992), however, suggest that interaction of agonists with the two conserved serines on helix V would alone be sufficient to induce the conformational changes necessary to trigger coupling to G proteins.

guidance. Extrapolating these findings to helix V with an agonist hydrogen bonded to Ser-204 and Ser-207 and the cationic amine orientated towards Asp-114 on helix 3 would produce a simulation in which the agonist is driven closer to Asp-79 on helix 2 of D_2. Since hydrogen bonds are easily broken, the agonist can reorientate to achieve a new equilibria. It is possible that only the bending action of helix V is necessary to achieve MODE 1, with MODE 2 following automatically and if the receptor is in the high affinity state, MODE 3 is achieved in combination with helix interface shear mechanism (see below). However, it should be mentioned immediately that middle prolines are also conserved in helices VI and VII. It seems reasonable therefore to suggest that the bending motions ascribed to helix V also occur in helices VI and VII. Hence, the *thrusting* action of helices V, VI and VII (there is a conserved serine/hydrogen bonding site on helix VII) are likely to work in unison pushing the agonist deeper into the binding pocket leading to the coupling of the receptor to intracellular G protein. It appears that evolution has provided *motors* to solve the problem of driving the agonist deep into the binding pocket.

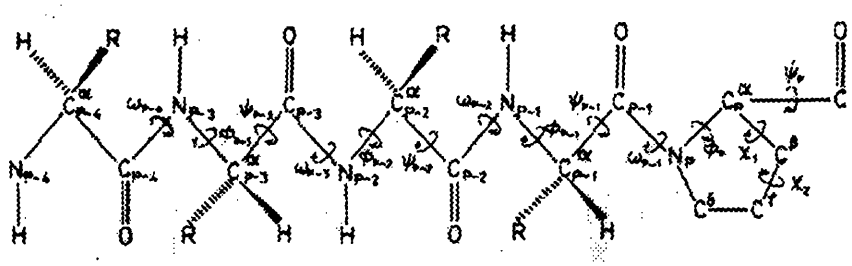


Figure 3.4.2; Conformational parameters of a polypeptide chain containing a middle proline. R = residue side-chain; ϕ, γ and ω are the classic back-bone torsion angles; χ_1 and χ_2 used here represent the internal torsion angles of the proline ring. Adapted from Sankaramakrishnan *et al.*, 1991.

Maloney-Huss and Lybrand (1992) only suggest that prolines introduce kinks into the helices and may function as joints in helices allowing more complex conformational shifts and reorientations - the relevance of the work of Sankararamakrishnan *et al.*, (1991) is not discussed. Similarly, Trumpp-Kallmeyer (1992) merely point out: "Interestingly, the

presence of a Pro residue in a TM α-helix has been studied experimentally and has been shown to have a hinge function, inducing oscillations of the two helical arms (Riegler, 1990)". Again no mention of the detailed findings of Sankararamakrishnan *et al.*, (1991).

3.4.3 Helix Interface Shear Mechanism.

Lesk and Chothia (1984) noted: *that many proteins undergo conformational changes in response to changes in state of ligation. The switch between specific conformational isomers is part of the mechanism of their function.* It is found in proteins that contain extensive domain-domain interfaces. This phenomena they called: *helix interface shear* mechanism which they described in terms of its role in domain closure in proteins. Essentially, helix interface shear mechanism is used by some proteins to allow individual helix rotations which are cumulative as they are transmitted from one helix to the next.

The possible role of the helix interface shear mechanism in the binding and release of ligands (in particular agonists) in GPCRs has never been properly discussed in the literature. Maloney-Huss and Lybrand (1992) alone have suggested that large scale helix-helix motions are necessary to permit entry and exit of ligands. However, the role of helix interface shear mechanism in allowing individual helix rotations to be cumulative as they are transmitted from one helix to the next was not appreciated by Maloney-Huss and Lybrand. Instead they refer to: "a general rotation or twist of the entire helix bundle". It is surprising that such a fundamental phenomena (i.e. helix interface shear mechanism) is not widely known or fully appreciated amongst GPCR modellers. This topic is considered in more detail in chapter 8.

4. Modelling The Dopamine G Coupled Receptor Proteins

4.1 Role of Three-Dimensional Structures

The belief of structural biology that function follows form had its origin in the treatise by Anfinsen (1959), *The Molecular Basis of Evolution*. Anfinsen remarked that "Protein chemists naturally feel that the most likely approach to the understanding of cellular behaviour lies in the study of structure and function of protein molecules." Summers & Karplus (1989) later noted that "A knowledge of the three-dimensional (3D) structure of a globular protein is an essential first step for the understanding of its biological function." Johnson *et al.* (1994) extended this fundamental truth to include the general class of integral membrane proteins. Fasman (1989) has observed that the achievement of protein crystallography over the past three decades lies in the fact that the structure and function of proteins is now often understood at the atomic level.

Unfortunately, whereas there are numerous globular proteins resolved to 2 Å or better, only one intermediate resolution structure of an integral membrane protein exists - photosynthetic reaction centre (PRC) from the prokaryote *Rhodopseudomonas viridis* (Deisenhofer *et al.*, 1984). This discrepancy in number of known structures owes much to the fact that integral membrane proteins are extremely difficult to crystallise; indeed, membrane proteins account for less than 1% of the protein structures available. In order to maintain the native structure, the lipid-protein interaction must be conserved during the crystallisation process. This poses enormous difficulties for workers aiming to obtain 3D X-ray crystal structures of integral membrane proteins. Consequently, crystallographic studies on membrane proteins have been considered to be *more art than science* (Garavito, 1990).

4.2 The Need For Protein Modelling

The catecholamine binding G protein coupled receptors (GPCRs) are integral membrane proteins and have never been crystallised though two-dimensional crystals of rhodopsin, which binds a chromophore ligand, have been used to produce a low resolution 2-dimensional projection map (Gebhard *et al.,* 1993). Obtaining experimental 3D structures of eukaryotic integral membrane proteins remains only a possibility in the medium to long term; the preponderant obstacles remain: obtaining "good" quality crystals - crystals of sufficient size and order such that they diffract X-rays to high resolution and the procurement isomorphous heavy atom derivatives. In this vacuum, the protein modelling community has been called upon to develop physically plausible 3D models - in particular: 3D models of channel membrane proteins and GPCRs. In the absence of a high resolution 3D structure, such modelling must remain largely speculative. However, such models are very useful in providing a theoretical framework for assessing evidence derived from multiple sequence alignments, biophysical experiments, site-directed and deletion mutagenesis, molecular dynamics studies and structure-activity relationships. Their value ultimately lies in their ability to aid the medicinal chemist in the rational drug design approach and molecular biologists in the design of new receptor engineering experiments (Humbler and Mizadegan, 1992).

More specifically, the value of accurate 3D models were implicitly emphasised in the IBC (1989) Symposium on Schizophrenia held in London, UK. It was noted in the meeting that drugs currently available for the treatment of schizophrenia have a rather "dirty" receptor profile and thus exhibit a range of disturbing side-effects by interactions (i.e. binding) at diverse sites including α_1-adrenoceptors, muscarinic cholinoceptors and histamine receptors. Hence, accurate models of the dopamine receptors (which are the targets for current drug treatments for schizophrenia and Parkinson's disease) can help medicinal chemists design ligands to bind specifically to the dopamine receptors. However, it was also noted that a greater understanding of the different sub-types of dopamine receptors is required due to the disturbing occurrence of side-effects related to interactions with the dopamine receptors (movement disorders caused by induced parkinsonism and tardive dyskinesia). Hence, models for each sub-type of dopamine receptor are required to help

medicinal chemists target drug treatments more effectively. Though our current understanding of the action of dopamine receptors has gaping holes - only by targeting specific dopamine receptors can medicinal chemists and pharmacologists observe which sub-type dopamine receptors are responsible for the observed side-effects. Indeed, the discovery of at least five dopamine receptor subtypes and their genes paves the way for new approaches to treatment of schizophrenia (Crow and Harrington, 1994). Hence it follows that accurate 3D models of each of the five DA receptor subtypes would aid medicinal chemists in the rational design of new highly selective neuroleptics (i.e. antagonist drugs) for the treatment of psychotic disorders.

4.3 Fundamentals of Structure Prediction

Protein modelling requires a range of skills and inputs both from experimental and theoretical workers. Molecular biologists must first obtain the primary structure since without detailed knowledge of the amino acid sequence it is not possible to proceed to the next step: structure prediction. This follows from the fundamental observations of Anfinsen *et al.* (1961) who demonstrated very elegantly that ribonuclease could be denatured and refolded *in vitro* without loss of enzymatic activity. A decade later, Anfinsen (1973) pointed out that he had been disturbed by some aspects of the 1961 series of experiments; in particular the fact that the successful refolding of the protein frequently took hours. Dintzis (1961) demonstrated that the time to synthesise a 124 amino acid protein with several disulphide bonds such as ribonuclease would take 1.5 minutes. Canfield and Anfinsen (1963) showed that egg white lysozyme was synthesised in less than 3 minutes. While these experiments gave no information concerning the formation of secondary or tertiary structure, Anfinsen clearly thought that *in vitro* refolding took much longer than might be expected (around 2 minutes to complement the time taken to synthesis a 124 amino acid protein). The difference between *in vitro* refolding and *in vivo* chain synthesis rates was explained by the work of Goldberger *et al.* (1963) who discovered an enzyme system in the endoplasmic reticulum which when added to solutions of reduced ribonuclease or to the protein containing randomised disulphide bonds, catalysed the rapid formation of the correct, native disulphide pairing in a period less than the requisite 2 minutes. Hence, protein tertiary structure is specified by the primary amino acid sequence

with regard to its usual environment. Rees (1990) succinctly puts Anfinsen's, albeit pioneering, work into realistic perspective: *Anfinsen's pioneering studies demonstrate that for certain proteins, the amino acid sequence determines the three-dimensional structure. For these cases, knowledge of the sequence is equivalent to knowledge of the structure.*

There have been numerous attempts to understand and exploit this phenomenon. Cohen *et al.* (1983) has grouped the approaches under two broad headings: (i) the direct use of energy-minimisation techniques and (ii) a two-step process *(the combinatorial approach)* that converts the sequence into a secondary-structure representation followed by the construction of a 3D structure by the packing together of the secondary structure elements. The somewhat heuristic combinatorial approach was developed by Fred Cohen (Cohen *et al.,* 1979). In reality, the combinatorial approach is only used where no known homologous protein structure exists. The third and most successful (but often least applicable) approach to structure is known as homology modelling (Blundell *et al.,* 1987). Homology modelling has been referred to as "comparative modelling" (Greer: 1981 and 1991). More recently the concept of homology modelling has been extended into a prediction strategy known as "knowledge-based" modelling - Blundell *et al.* (1987) and Johnson, *et al.* (1994). Given the importance attached to these three approaches to predicting protein structure, each will now be discussed in some detail.

4.3.1 Direct Use of Energy Minimisation Techniques

The energy of a molecule in the ground electronic state is a function of its atom positions. The Born-Oppenheimer surface is the multidimensional surface that describes the energy of a molecule in terms of atom positions. In molecular mechanics (force field calculations) it is often referred to as the potential energy surface. This topic is considered in more detail in a later section concerned with energy minimisation strategies so force field function theory will not be considered in detail here. However, powerful computers are able to calculate the energy of even a large molecule such as a protein in seconds - i.e. based on a static set of x, y and z co-ordinates of its atoms. While a 2D or even 3D Ramachandran plot corresponding to the rotation of a single bond in a protein structure will take a little longer - the 3D representation of the Ramachandran plot displays gullies,

valleys and plateaus. The 3D energy surface is multi-dimensional and to calculate the surface representing a large protein (including all back-bond ϕ and ψ torsion angles plus all possible χ side-chain torsion angles) is all but impossible using current day supercomputers.

This problem even extends to real proteins since if proteins adopted their 3D structure by sampling every possible conformational space - proteins would not be able to achieve their 3D structure on any reasonable time scale. For example, a relatively small protein of 50 amino acids could adopt approximately 10^{50} different conformations (Sternberg and Thornton, 1978). Nevertheless, as Šali *et al.* (1994) point out: protein sequences do fold into unique native states in seconds (the Levinthal paradox[1] - Levinthal, 1969).

Energy calculations aimed at predicting the tertiary structure from just the primary amino acid sequence assume that the native structure represents the global minimum. Searching for the global minimum using empirical force fields has up to quite recently been considered not feasible. Not only is the computer time required excessive, but the method involves too many approximations and energy surface is extremely complex involving numerous minima. However, Šali *et al.* (1994) have demonstrated that the problem is much simplified if energy calculations avoid starting with a random coil and instead start with a semi-compact globule. The Levinthal paradox is resolved for polypeptides (~27 amino-acids in length); Šali *et al.* (1994) calculated that the number of possible conformations that need to be searched (and their energies calculated and ranked) is reduced from ~10^{16} (for random coil) to ~10^{10} which in turn lead to ~10^{3} transition states and then 1 native state (the global minimum) - figure 4.3.1. Hence, the size of the search for the native state is greatly reduced when the chain is semi-compact, as it is in real proteins (Dill, 1985). However, these new findings are based on a 27-bead chain and so side-chain torsion angles were not simulated. Also, the method is limited to chains which do not get trapped in local minima, i.e. the native state of the model has a pronounced global minimum on the potential surface. However, the results suggest that for small proteins, effort may be better applied to the derivation of a suitable potential function rather than the design of folding algorithms to predict tertiary structure.

[1] There is not enough time for a protein molecule to sample all possible conformational states to eventually locate the biologically active state - this is the Levinthal paradox.

For larger proteins, the vastness of the conformational space argues in favour of a what has become known as the protein folding problem or pathway. The native structure will be at the minimum free energy of the kinetically accessible conformations. This minimum need not be the global minimum. The folding pathway only samples a much reduced subset of all possible conformations. It is thought that stretches of amino acids (up to 18 amino acids in length - Wetlaufer, 1973) act as nucleation sites, around which the remainder of the protein folds thereby reducing the number of conformations searched. Possible nucleation sites are the α-helix and β-strand. Designing folding algorithms to predict final tertiary structure rather than sampling all possible conformations and risk getting trapped in local minima (the multi-minima problem) offers some hope of predicting tertiary structure from knowledge of the amino-acid sequence.

In another approach to solve the Levinthal paradox (Shakhnovich, 1994; also see: Maddox, 1994), Shakhnovich designed sequences of an 80-monomer protein which provided very low energy in the target (i.e. *native*) structure. The designed sequence was then subjected to lattice Monte Carlo simulation of folding. In each run, the model protein folded from random coil to the unique native conformation without encountering metastable states en route thereby effectively solving the multiple minima problem. Shakhnovich's results suggest that is the thermodynamically orientated selection of sequences which makes the native conformation a pronounced deep minimum of energy and in turn solves the problem of kinetic accessibility of this conformation as well. However, Shakhnovich did predict that 5% of the protein would not fold correctly and claimed that there is a considerable body of unpublished calculations to support his conclusions.

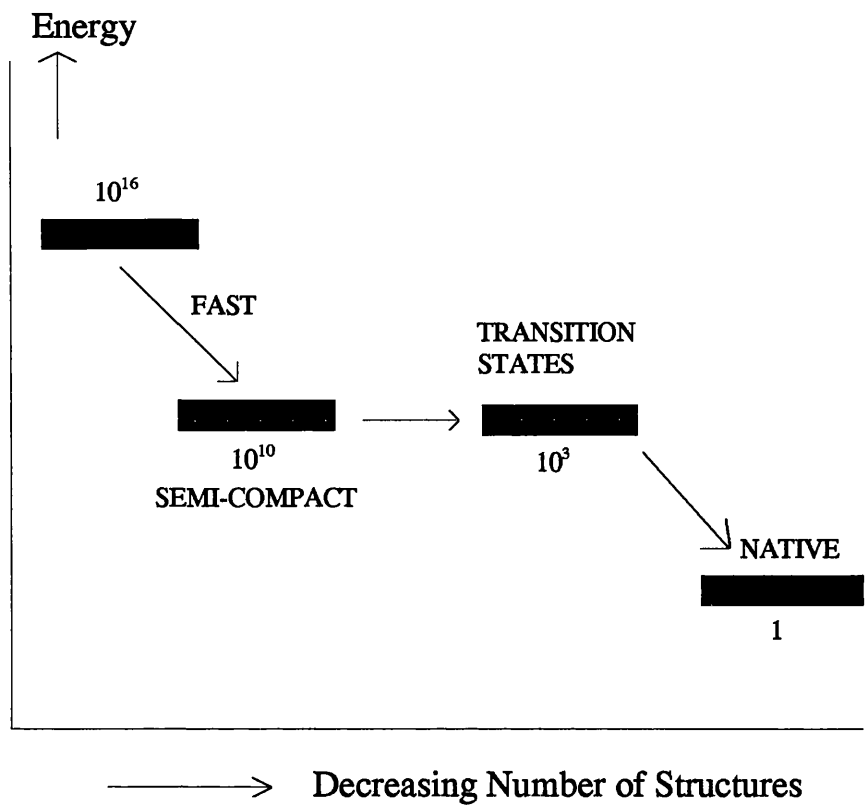


Figure 4.3.1 Folding of a polypeptide chain of just 27 amino acids (adapted from Šall *et al.*, 1994). Protein folding starts with a rapid collapse from a random-coil state to a random semi-compact globule. A slow, rate-determining search follows through the semi-compact states to find a transition state from which the chain folds rapidly to the native state. The reduced number of conformations that need to be searched in the semi-compact globule (~10^{10} as against 10^{16} for the random coil state) leads to the resolution of the Levinthal paradox. A protein needs to fold in seconds (possibly 2 or 3 minutes) which would be an impossible task if the protein had to sample every possible conformational state associated with the random-coil state (the Levinthal paradox). Šall, Shakhnovich and Karplus (1994) success in resolving the Levinthal paradox (albeit for a small protein of just 27 amino acids and ignoring side-chains) suggests that the bottleneck in structure prediction may be solved using the derivation of a suitable potential function rather than the design of folding algorithms. Karplus's group will undoubtedly endeavour to improve on their current findings to resolve the structure of larger proteins. Indeed, Shakhnovich (1994) has designed sequences of 80-monomers which folded from random coil to the target (i.e. *native*) structure in just 6 million time steps (see main text).

4.3.2 Homology Modelling

Zuckerkandl and Pauling (1965) suggested that amino acid sequence data could be used to chart evolution among homologous proteins. Greer (1981) noted that many diverse proteins have been classified into families on the basis of sequence homology. The similarity of three-dimensional structure between known homologous proteins suggested it should be possible to model other members of the same family by comparative model building. Greer showed that it was relatively easy to model homologous regions which he applied to: mammalian proteases (Greer, 1981, 1990a) and the design of novel renin inhibitors (Greer, 1990b). Greer (1981) noted that in the case of variable regions detailed comparisons with known structures on the basis of length and residue character allowed structurally variable regions to be modelled by homology and the necessity to build from intuition or from energy considerations was greatly reduced.

Knowledge-based modelling[2] (comprehensively reviewed by Johnson *et al,* 1994) depends on analogies between a protein that is to be modelled and other proteins of known three-dimensional structure at all levels in the hierarchy of protein organisation: secondary structure, motifs, domains and quaternary or ligand interactions (Fasman, 1989). Homology modelling purely relies on the occurrence of in the unknown protein sequence homologies to known sequences, structures and fragments of structures that have been solved through X-ray crystallography or NMR (for small proteins). Some authors suggest that the homology modelling approach requires sequence similarity of approximately 50% (e.g. Wishart and Muir, 1990) or identity > 25% (Rost and Sander, 1993). However, Wishart and Muir (1990) point out that it is: "relatively rare to find attempts at homology modelling where sequence identity is less than 70%". For this reason, homology modelling has very limited application. In reality, hybrid approaches are frequently used depending on the protein being modelled. For example, Wishart and Muir (1990) adopted a "hook or by crook" approach based on the hierarchical approach of Cohen *et al.* (1979, 1980) and the homology approach of Blundell (1987) to predict and model the structures of: mandelate racemase, cellobiohydrolase, dehalogenase and defensin HNP-3.

[2] COMPOSER is a mini expert system which exploits the knowledge-based modelling approach and has been used, for example, to model human plasma kallikrein and human neutrophil defensin (Johnson *et al.*, 1990).

The number of cloned sequences now exceeds 50,000 in number (Johnson *et al.*, 1994) and is growing at an exponential rate far outstripping the rate at which protein structures are being resolved. Wishart and Muir (1990) are of the opinion that the pursuit of the *Holy Grail* - the prediction of 3D protein structure using only the knowledge of the amino acid sequence - may soon no longer be a goal beyond our reach but one within our grasp. The brilliant work of Shakhnovich (1994) supports Wishart and Muir's optimistic view. It should be noted however that Shakhnovich's approach involved using a lattice to perform a Monte Carlo simulation of folding and therefore does not rely on searching a database of known structures to locate regions of homology.

The number of different protein folds adopted by globular proteins is estimated by Blundell and Johnson (1993) to be in the range 500 to 700 and Chothia (1993) suggests a figure of ~1000 is more reasonable. Only 50% may be known. Protein crystallographers and NMR spectroscopists tend to select similar proteins that are amenable to their techniques (Johnson *et al.*, 1994) which suggests that 50% may be an overestimate. However, Blundell and Johnson (1993) argue that: "we should move toward an experimental definition of one example of each common fold" and add: "if methods to identify the folds from their sequences can be developed and if comparative modelling can be extended to more distantly related protein topologies, then we should be able to provide at least rough indications for most sequences as they become available."

Shakhnovich's *ab initio* approach is not reliant on a database of known proteins with different folds. Also, Shakhnovich's approach is likely to see considerable improvements in potency of several orders of magnitude. Should Shakhnovich's algorithm be developed to the point that it is released as a mini-expert system it could be used to predict the structure of any cloned sequence. In contrast mini-expert systems such as COMPOSER remain hampered by the limited number of known protein structures. Shakhnovich's approach may be exploited to predict as yet unseen protein folds. Should this happen it would herald a new era in protein structure prediction.

4.3.3 The Combinatorial Approach

As Rost *et al.* (1994) points out: the 3D structure of a new sequence can be predicted from the sequence fairly accurately if a homologue with significant sequence similarity exists in Brookhaven data bank of experimentally determined structures. However, for more than 80% of sequenced proteins there is no homologue of known 3D structure. For these proteins Rost also suggests that the way out is to reduce the problem to a simpler one that is amenable to a partial solution, i.e. the starting point should be protein secondary structure prediction using the known primary amino acid sequence. The next logical step would be to construct the 3D structure from the predicted secondary structures. Hence, the combinatorial approach offers the opportunity to model proteins which lack homology with any know homologue. The DA receptors have been cloned and sequenced but they lack homology with any known 3D structure. We will now consider the first stage of the combinatorial approach: secondary structure prediction and assess its value in helping to solve the GPCR protein modelling problem.

4.3.3.1 Secondary Structure Prediction

The most popular methods are frequently compiled in the form of a suite of programs designed to be of general use to molecular biologists and protein modellers. For example, the Wisconsin Genetics Computer Group package (Devereux *et al.,* 1984) includes a suite of protein prediction algorithms which are frequently used by protein modellers to analyse the primary structure of an unknown protein (i.e. unknown in the sense that the structure has not been determined experimentally). Hence, we must begin with a survey of protein structure prediction algorithms and critically examine their potential role in aiding the GPCR modelling process.

4.3.3.2 Secondary Structure Prediction As Applied to GPCRs

Tertiary protein structure consists of three types of secondary structure (Q_3). There is the α-helix secondary structure first predicted by Pauling *et al.* (1951). The second

type of secondary structure is the β-sheet. The third type of secondary structure is the random coil.

While the three types of secondary structure are found in membrane proteins, the rules for protein folding of the membrane component of integral membrane proteins such as GPCRs are clearly different from those of water-soluble globular proteins. The single biggest difference is that extended random coils have not been found in the membrane component of integral membrane proteins. Transmembrane α-helices are found in PRC and bR, while β-sheets forming β-barrels are believed to form the membrane component of channel pores. Because of the need to satisfy hydrogen bonding, random coils can not exist in the bilayer as there are no water molecules to satisfy the hydrogen bonding requirements of polar side-chains or polar atoms of the protein backbone. Hence, deriving secondary structure predictions from GPCR sequences usually starts with a search for apolar segments long enough to span the lipid bilayer (typically 20 to 30 amino acids long - length depends on tilt of helix relative to bilayer; Engelman *et al.*, 1986). Hydropathy plots (considered in detail in chapter 5) have proved very useful in identifying putative TM segments of GPCRs (reviewed by: Fasman and Gilbert, 1990; also: Jähnig, 1990)

4.3.3.3 The Chou-Fasman Prediction Methods

The Chou-Fasman algorithm for the prediction of protein structure is one of the most frequently cited methods in the literature (Fasman and Gilbert, 1990). The explanation for this lies in its relative simplicity and its reasonable high degree of accuracy (Prevelige Jr. and Fasman, 1989) and can be applied without the use of a computer (Sternberg and Thornton, 1978). The method was not made available by the authors in the form of a computer program and consequently there are several published computer versions of this popular method in use.

The x-ray determined structures of 15 proteins containing 2473 amino acid residues were analysed. The frequency of occurrences of an amino acid type in a α helix, β-sheet and loop was calculated. This information was in turn used to estimate conformational

parameters (table 4.3.3.3) for each residue by considering the relative occurrence of a given residue within a protein, its occurrence in a given type of secondary structure, and the fraction of amino acids occurring in that type of structure (Chou and Fasman, 1974a). These parameters, symbolised by P_α , P_β and P_c, respectively, probably hold data concerning the physical-chemical parameters which characterise protein stability. Hence the method falls into the statistical approach to secondary structure prediction.

The next step required the formulation of rules to help predict secondary structure and thereby make maximum use of the calculated parameters: P_α, P_β and P_c (Chou and Fasman, 1974b):

1. A helix region is defined by a cluster of four helical residues (H_α or h_α) which signifies helix nucleation - starts at the N-terminus. The helix is then propagated towards the C-terminus until the tetrapeptide window drops below 1.00 (or 100 if the parameters are multiplied by 100). Any segment which is at least 6 residues in length with a average $P_\alpha > 1.03$ and average $P_\alpha >$ average P_β is predicted as helical. A further caveat is that Pro cannot occur in the inner helix or at the C-terminal end but can occur at position 1 at the N-terminus where it is the second most common residue occupying this position (Richardson and Richardson, 1989). Consequently, given that Pro residues are easily tolerated in the middle part of TM helices in bacteriorhodopsin (Henderson *et al.*, 1990) this renders the Chou-Fasman method completely inappropriate for predicting the secondary structure of the membrane components of membrane bound proteins. If experimental evidence suggests the membrane component is essentially helical and that it contains conserved Pro residues, the Chou-Fasman method should not be used for secondary structure prediction.

2. A β region is defined by a cluster of three β formers or three β formers in a sliding window (no gaps allowed) along the primary structure. The β sheet is generated in both directions until terminated by a tetrapeptide window has an average $P_\beta < 1.00$. Any chain (i.e. segment) with an average $P_\beta > 1.05$ and greater than the average P_α is presumed to be a β sheet. Unfortunately, β-

branched side-chains such as Ile, Val and Thr are regarded by Chou-Fasman as strong β formers. However, these residues are well represented in the putative TM helices of GPCRs and hence any β-sheet prediction obtained using this method must immediately be suspect.

Chou and Fasman (1977) reported a prediction method for the chain reversal regions (i.e. β-turns) of globular proteins. Keeping to their statistical approach Chou and Fasman calculated a conformational parameter (P_t) defining the probability of a β-turn starting at residue *i*. However, positional preferences for particular residues in the β-turn were found. In particular, Pro had a strong preference for *i*+1 (30%) whereas at *i*+3 the frequency of occurrence of Pro was only 4%. Chou and Fasman (1979) reported an automated computer prediction method which took into account position preferences in the β-turn. Bend frequencies for each residue were calculated for each position in the bend (f_i , f_{i+1} , f_{i+2} , f_{i+3}) and p_t calculated:

$$p_t = f_i * f_{i+1} * f_{i+2} * f_{i+3}$$

A cutoff value of $p_t > 0.75 * 10^{-4}$ is used to predict β turns. Given that β-turns in GPCRs are likely to occur at helix ends it is possible that predicted β-turns in the primary structures of GPCRs would help locate helix end positions. Indeed, as Beverley Green (1990) points out in her review of structure prediction methods suitable for membrane proteins (considered in section: 4.3.3.7.5): many workers have implied that turn predictions can be utilised to locate ends of TM helices (Paul and Rosenbusch, 1985; Wilmot and Thornton, 1988; Shriver *et al.*, 1989).

Helical Residues	P_α	Helical Assignment	β-Sheet Residues	P_β	β-Sheet Assignment
GLU	1.53	H_α	MET	1.67	H_β
ALA	1.45	H_α	VAL	1.65	H_β
LEU	1.34	H_α	ILE	1.60	H_β
HIS	1.24	h_α	CYS	1.30	h_β
MET	1.20	h_α	TYR	1.29	h_β
GLN	1.17	h_α	PHE	1.28	h_β
TRP	1.14	h_α	GLN	1.23	h_β
VAL	1.14	h_α	LEU	1.22	h_β
PHE	1.12	h_α	THR	1.20	h_β
LYS	1.07	I_α	TRP	1.19	h_β
ILE	1.00	I_α	ALA	0.97	I_β
ASP	0.98	i_α	ARG	0.90	i_β
THR	0.82	i_α	GLY	0.81	i_β
SER	0.79	i_α	ASP	0.80	i_β
ARG	0.79	i_α	LYS	0.74	b_β
CYS	0.77	i_α	SER	0.72	b_β
ASN	0.73	b_α	HIS	0.71	b_β
TYR	0.61	b_α	ASN	0.65	b_β
PRO	0.59	B_α	PRO	0.62	b_β
GLY	0.53	B_α	GLU	0.26	B_β

Table 4.3.3.3; Conformational parameters: P_α and P_β used to describe the propensity for amino acids to be found in a helix and β-sheet. Helical assignment descriptors: H_α = powerful α-helix former, h_α = α-helix former, I_α = feeble α-helix former, i_α = apathetic α-helix former, b_α = α-helix breaker, B_α = powerful α-helix breaker. β–sheet assignments: H_β = powerful β former, h_β = β former, I_β = feeble β former, i_β = apathetic β former, b_β = β breaker, B_β = powerful β breaker. Chou and Fasman (1974b)

4.3.3.4 The GOR method for predicting secondary structure

The Garnier, Osguthorpe and Robson (GOR) method for predicting the secondary structure of globular proteins exploits Information Theory, which has found wide use in many applications such as monitoring telecommunication networks. The ground work for the GOR method was done by Pain and Robson (1970) and Robson and Pain (1971) who recognised that a sequence of amino acids could be treated as a message decoded and converted by the folding mechanism into another message consisting of a sequence of conformational states. Essentially the GOR prediction method considers quantitatively the disposition of each type of residue, individually and as part of a pattern, to adopt different conformations.

It can be demonstrated that the information brought by a complex event can be decomposed into a sum of information brought by more simple events. Some of the information brought by these simple events might overlap, so it is important to recognise overlapping regions and avoid counting the same information twice. Hence the conformation that a single residue might adopt will depend not only on its single residue preference for a given conformation, but will also depend on such factors as: interactions inside each short-chain region, middle-range interactions between chain regions adjacent along the chain and long-range interactions between different chain regions that happen to be close to one another in the 3D structure (Ptitsyn and Finkelstein, 1983); solvent accessibility (water for globular proteins plus lipid for membrane proteins). To take account as much as possible of the many factors governing the conformation of a single residue it is necessary to consider single residue information, pair residue information, triplet information etc. However, a problem immediately arises here. Table 4.3.3.4 clearly shows that for four conformational states[3] (H, helix; E, β-sheet; C, coil; T, β-turn) there are 80 entries (and hence 80 separate probabilities to calculate) using single residue information whereas for quadret information there are 640,000 entries - this would require a massive database of residues in known structures.

[3] In reality, it should be possible to just calculate any three sets of probabitlities coresponding to just three conformational states which would leave the probability values coresponding to the "missing" set.

The GOR method uses directional information values (Garnier *et al.*, 1978) which were calculated using a database of just 26 proteins with approximately 4,500 residues. Gibrat (1986), using a larger database of 75 proteins with 12,757 residues recalculated the directional information values to establish the GOR II method. In the final GOR III version, pair-information was used to greater effect instead of directional information as in the GOR II method (Garnier and Robson, 1989). Consequently, secondary structure prediction of globular proteins has improved from approximately 55% to 65% for three states (Q_3: helix, sheet and coil).

How applicable is the GOR III method to predicting secondary structure in the membrane component of integral membrane proteins? GOR III (or GOR or GOR II) should not be used to predict secondary structure in membrane proteins. The reason is quite simple: the databases used to calculate probabilities values is based on structures of globular proteins. This issue is considered in more detail in section 4.3.3.6.

Information Type	Number of Entries
simple (single) residue information	20 * 4 = 80
pair information	20 * 20 * 4 = 1,600
triplet information	20 * 20 * 20 * 4 = 32,000
quartet information	20 * 20 * 20 * 20 * 4 = 640,000

Table 4.3.3.4 Residue information coresponding to the 4 conformational states.

4.3.3.4.1 Fixing the intrinsic problem with the GOR method

The GOR method has a natural limitation, by itself it fails to take account of the influence of long-range interactions on secondary structure. The GOR method takes account of local (as well as middle-range interactions) but to take account of the effect of long-range interactions requires some input concerning the likely tertiary structure. Indeed, Gibrat *et al.* (1991) postulated that algorithms (e.g. such as the GOR method) are prone to predict secondary structure to a maximum accuracy of around 65% because of the limited influence of short-range interactions.

The majority of proteins belong to a given class (α, β, $\alpha + \beta$ or α/β; Richardson, 1981) and have a very limited set of typical topologies. For example, the recent elucidation of human chorionic gonadotropin (HCG) reveals that it is clearly in the β class (Lapthorn *et al.*, 1994). Successful prediction of class of protein allows secondary prediction algorithms to take account of long-range interactions (Ptitsyn and Finkelstein, 1983)[4][5]. Long-range interactions for each chain region can be modelled by the interaction of each chain with an averaged hydrophobic template (Lim, 1974a,b). The use of stereochemical theory of globular protein secondary structure to take account of long-range interactions was first applied by Schiffer and Edmundson (1967).

Ptitsyn and Finkelstein (1983) reminds us that while the stereochemical method has the benefit of being *a priori,* it does not lead to more correct secondary structure predictions. These workers have developed a physical theory of protein secondary structure which takes account of both local interactions inside each chain region and long-range interactions between different regions (and by default middle-range interactions are also included). Their model uses stereochemical theory to evaluate local interactions and thereby judge the relative stabilities of α-helices and β-structures for different amino acids in synthetic

[4] Geisow and Roberts (1980) have observed that the residue preferences for secondary structure vary with the protein class.

[5] Garnier *et al.* (1978) in their key paper describing the GOR method noted that if the type of protein (i.e. class) can be determined even approximately by circular dichronism (or any other suitable prelimary prediction) this data could be used to bias the method. In this way, 57% of residue states can be correctly predicted.

polypeptides. Long-range effects are modelled by interacting each chain region with an averaged hydrophobic template.

Both strategies have been fully computerised and the algorithm has been successfully applied in blind predictions (made before the x-ray structure becomes available). The algorithm accurately predicted the α-helices and β-strands in uteroglobin and Tyr-tRNA-synthetase. However, the algorithm has difficulty breaking the 65% secondary structure accuracy barrier even though long-range interactions are incorporated in the method. Also, the algorithm has not been successfully applied to membrane bound proteins. The reason for this limitation is tied up with the preference for α-helical structures to dominate in the membrane part of the protein as these structures are very successful at satisfying the requirement for internal hydrogen bonding and so single TM helical structures are favoured in single membrane spanning virus coat proteins, e.g. M13 coat protein.

4.3.3.5 Profile Network from Heidelberg - PHD

The prediction of protein secondary structure at better than 70% accuracy using information about a structure contained in a multiple sequence alignment has been successfully achieved (Rost and Sander, 1993; Rost *et al.*, 1994). A protein sequence contains information about (Paul Emsly, *personal communication)*:

- spacial information
- active site (enzyme)
- binding site (receptor proteins)
- protein-protein interaction
- membrane transport
    -
        -
            - + mutational noise

By using a family of primary structures to remove "noise" and increase the level of "spacial information" PHD increases secondary structure prediction reliability by 6 to 8%. This increase in reliability is somewhat analogous to the improvement claimed for the probabilistic method of Garnier *et al.* (1978) where the information available from a family of sequences was used to improve secondary structure prediction. Zvelebil *et al.* (1987) applied the GOR method to nine families of homologous proteins and claimed a 9% improvement for secondary structure prediction. This was achieved by simply allowing the GOR method to favour the prediction of loops and coils in regions with high sequence variation.

Various pattern recognition problems have lead to widespread use of neural networks in general and "feed-forward" networks in particular. PHD consists of 3 layers (or levels): 2 network layers and 1 layer averaging over independently trained networks. The first level is a *sequence-to structure* net, which classifies strings of adjacent residues (=sequence pattern) into the 3 secondary structure classes (Q_3): α-helix (H), β-strand (E) and loop (L). Probability for each state (Q3) for the central residue is calculated; 010 would imply a zero probability for H, an absolute probability for E and a zero probability for L. A sliding window methodology is used and probabilities are calculated for each central residue in each possible triplet in each window. The frequency of occurrence of each of the 20 amino acids at one position in the multiple sequence alignment is calculated and used aid the classification of the central residue. The output (units for α-helix (H), β-strand (E) and loop (L) acts as input to the second level.

In level one there no account is made of the fact that consecutive patterns are correlated, e.g. for a helix consisting of at least 3 consecutive patterns. The second level performs this correlation and thereby improves the probability calculation (again for the central residue) for α-helix (H), β-strand (E) and loop (L).

The third level acts like a jury. It takes the calculated probabilities for each central residue (calculated by independently trained networks operating at level 2) and simply takes the highest probability level and outputs that one. The well known catch phrase *winner takes*

all has been used to describe the role of level 3. In one sentence, PHD is able to classify patterns according to their intrinsic correlation (common information).

4.3.3.5.1 Why PHD is the best secondary structure prediction algorithm so far for non-membrane proteins

PHD is better in 4 ways:

1. the overall accuracy (70.8% for globular water-soluble proteins) is greater than any other method (4 to 6% improvement).

2. β-strand per-residue accuracy is 65.4% (best of the rest is GORIII's 46%).

3. the length of predicted secondary structure elements is more like real proteins than that for other typical prediction methods.

4. residue's secondary structure is predicted with greater probability (20% of residues have a reliability probability > 90% and more than 50% of residues have a reliability probability > 82%).

In addition, PHD predicts the content of secondary structure with less than 10% error. PHD is currently available free of charge by means emailing a sequence to PHD@EMBL-Heidelberg.DE. The mail server at EMBL forwards the sequence directly to the feed forward neural network and a reply containing full secondary structure prediction information (and secondary structure content) is generally returned within 4 hours (sometimes within 2 hours). At the moment, PHD is the best secondary prediction algorithm available for the above reasons. However, the ultimate goal is reliable prediction of tertiary (3D) structure, not 100% single residue accuracy for secondary structure (Rost *et al.*, 1994).

PHD's overall accuracy (70.8% for globular water-soluble proteins) is especially significant in the light of Gibrat *et al.* (1991) postulation that the prediction of secondary structure may be limited to approximately 65% because of the limited influence of short-range interactions. They pointed out that local amino acids may contribute on average 65% to the torsion angles (φ and ψ) adopted by residues in proteins which in turn decide local secondary structure[6]. However, they add a caveat: short range forces may carry more influence on local torsion angles but that the prediction methods so far (i.e. up until 1991) have been unable to extract this extra information. If this is the case, it would appear PHD's overall accuracy of 70.8% has managed to extract at least some of this extra information. Rost and Sander (1993, 1994) did note that the use of multiple sequence alignments of homologous structures significantly improved PHD's accuracy. So it would seem that PHD has broken the 65% prediction accuracy barrier and Rost and Sander deserve proper acknowledgement for this remarkable achievement.

4.3.3.5.2 How good is PHD at predicting secondary structure of membrane proteins?

Rost and Sander (1993) acknowledge that membrane proteins have a different physical environment from water-soluble globular proteins and, hence, different rules *have to be learnt* to correctly predict secondary structure. Put another way, the rules for protein folding are different for membrane proteins. These workers used PHD to see how accurately the four chains of the membrane protein photosynthetic reaction centre (1prc_C, 1prc_H, 1prc_L and 1prc_M) are predicted. The prediction was, as expected, below that for water-soluble globular proteins. Their results showed that for PHD the claims of accuracy made for PHD only apply to water-soluble globular proteins. β-sheet tends to be overpredicted at the expense of α-helix predictions. This is in line with the earlier observations of Jähnig (1989). He observed that all membrane proteins, or their membrane-incorporated parts, are predicted to be in β-strand conformation. For water-soluble globular proteins, β-strand conformation is highest for Val, Ile, Tyr and Leu. These residues are also well represented in putative TM α-helices (nearly 50% of TM amino acids [Deber *et al.*, 1986]; ~40% of TM residue composition [Deber *et al.*, 1992]) and

[6] Secondary structure prediction algorithms are based on short-range interactions, i.e. they use the information drawn from the local amino acid sequence.

consequently are predicted as forming parts of β-strands. As Li and Deber (1993) point out:

> *While the relationships between amino acid sequence and structure of the extracellular and cytoplasmic domains of membrane proteins may follow the same rules that govern globular (soluble) protein structure, the helical structures of their transmembrane domain(s), which contain a preponderance of helix-destabilising (e.g. Val, Ile, Thr, and Gly) residues, cannot be immediately appreciated form these rules.*

In addition, Li and Deber (1993) note that their overall findings suggest that the α-helix is the natural choice of conformation for a peptide of around 20 amino acids in length in a membrane environment. Further, that a helical conformation will arise "automatically" in a peptide above a threshold hydrophobicity.

The driving force favouring α-helical conformation in within membranes is the requirement for maximal hydrogen bonding. Transfer of a hydrogen bonded C=O and N—H pair in a protein backbone from water into a non-polar environment has been estimated to have a favourable free energy ΔG_{trans} of -1.4 kcal mol^{-1} (-5.9 kJ mol^{-1}) compared to an unfavourable change of 4.1 kcal mol^{-1} (17.1 kJ mol^{-1}) for a non-hydrogen bonded pair (Martonosi, 1985).

4.3.3.6 Why Are Popular Secondary Prediction Algorithms So Poor When Applied To Membrane Proteins?

Some of the early developments leading to secondary prediction the popular secondary structure prediction methods we see today involved the correlation of the protein secondary structure with the amino acid composition. Szent-Györgyi and Cohen (1957) found that proteins with a high percentage of proline distributed throughout the sequence have low helical content using optical rotatory dispersion (ORD). In turn, Davies (1964) used ORD to show that a qualitative relationship exists between the helicity of a protein and the total percentage of those residues in a protein classified as helix breakers: Ser, Thr,

Val, Ile by Blout *et al.* (1960, 1962). Goldsack (1969) confirmed that the total content of Pro, Ser and Thr decreased the helicity in a globular protein. Given that all of these residue are well represented in the TM region of GPCRs which are characterised by a heptahelical motif and also in single span TM proteins such as virus coat proteins, it is clear that residues which are regarded as helix breakers in globular proteins are tolerated in the helical regions of the membrane component of membrane proteins. For example, the M subunit of the photosynthetic reaction centre of *R. viridis* has a proline near the middle of the third TM helix (Deisenhofer *et al.,* 1985), which provides experimental evidence that a proline does not necessarily represent a breaking point of a TM α-helix.

Indeed, Wallace *et al.* (1986) have evaluated the validity of using methods designed to predict the secondary structure of globular proteins (Chou and Fasman, 1974a,b; Garnier *et al.,* 1978) for predicting the secondary structure of membrane proteins. They concluded that these methods are inappropriate for predicting the secondary structure of membrane bound proteins (15 examined). Only two of these membrane proteins, crambin and the photosynthetic reaction centre from *R. viridis* had been determined by x-ray crystallography. The other membrane protein conformations were determined by physical-chemical techniques and so the these structures may have introduced errors into their study. However, it is clear that all of the popular methods of secondary structure prediction (including PHD) are not able to accurately predict the secondary structure of membrane proteins.

So why are popular secondary prediction algorithms so poor when applied to membrane proteins? The obvious answer to this question is that the secondary structure propensities of key residues differ in water and membrane environments. The first definitive experimental evidence for this came from Li and Deber (1992a,b) who synthesised a series of model 20-residue peptides with the hydrophobic segments buried in *N*- and *C*-terminal hydrophilic matrices. The prototypic sequence being: NH_2-(Ser-Lys)$_2$-Ala^5-Leu^6-**x**^{7}-Ala^8-Leu^9-**y**^{10}-Trp^{11}Ala^{12}-Leu^{13}-**z**^{14}-(Lys-Ser)$_2$-OH. x, y and z varied from x = y = z = Ala (identified as peptide 3A) to x = y = z = Gly (identified as 3G); see table 4.3.3.6. The object being to experimentally examine the likely role of helix-breaker Gly in TM helices. They noted that Gly (and β-strand promoters: Ile, Val and Thr - all β-branched residues)

are well represented in TM segments, for example: the TM segment of bacteriorhodopsin (103 to 130): TILAIVGADGLMIGTGLVGALALTKV (single letter code). Hence the secondary structure propensities of these residues in membrane environments may differ from those calculated by such workers as Chou and Fasman (1974b) which are based on a data-base of water-soluble globular proteins. Using circular dichroism of preparations of these peptides in water, in a membrane-mimetic [sodium dodecylsulfate (SDS)] medium, and in methanol they found that despite its backbone flexibility, Gly can be accommodated as readily as Ala into a hydrophobic α-helix in a membrane environment (table 4.3.3.6).

PEPTIDE	HELICITY Aq. BUFFER %	HELICITY 10 mM SDS %
SKSKALAALAWALAKSKSKS 12345678901234567890 A A A	30	100
SKSKALGALGWALGKSKSKS 12345678901234567890 G G G	2	92

Table 4.3.3.6 Synthesised peptides; one type (3A) with A (Ala) occuppying positions 7, 10 and 14; second type (3G) with G (Gly) occuppying the same positions. Helicity was determined by taking ellpiticity at 222 nm as a direct measure of peptide helicity (Engel *et al.*, 1991) and helicity of 3A in 10 mM SDS (Sodium dodecylsulfate - mimics membrane environment) is taken here to be 100%. Chou and Fasman (1974a, 1978) conformational parameter for helix (P_α) (based on a data-base of water-soluble globular proteins) of Ala and Gly is: 1.39 and 0.63 respectively. Hence from P_α values would predict that 3G would not form a helix since Gly is considered to be a helix-breaker. However helical propensity, as this set of results clearly demonstrate, depends on environment. Glycine residues clearly support peptide helicity in membrane environments. Single letter amino acid code: A, Ala; S, Ser; K, Lys; Leu, L; G, Gly; W, Trp; I, Ile; V, Val. Adapted from Shun-Cheng Li and Charles M. Deber (1992b).

4.3.3.7 Algorithms specifically designed or used to predict secondary structure of integral membrane proteins.

Argos *et al.*, (1982) developed a prediction algorithm for membrane-bound proteins using the physical characteristics of the 20 amino acids in conjunction with the postulated structure of bacteriorhodopsin. A hierarchic ranking of the 20 amino acids was compiled with regard to their preference to interact with the lipid bilayer and this was used to delineate likely membrane-buried regions in the primary structure. A helical wheel analysis could then be applied to determine which face of each helix faced the interior of the protein and which faced the surrounding lipid bilayer. The main problem with this method is that it is based on a proposed bacteriorhodopsin structure and hence the technique is rather obsolete given that a model of the structure for bacteriorhodopsin based on high resolution electron cryo-microscopy now exists (Henderson *et al.*, 1990).

4.3.3.7.1 Simple Hydropathy Schemes

Jähnig (1990) reiterated the fundamental question: "How can membrane-spanning helices or strands be predicted from the amino acid sequence?" One solution is the application of the simple hydropathy plot proposed by Kyte and Doolittle (1982). If the Kyte and Doolittle algorithm is applied to bacteriorhodopsin six hydrophobic segments are clearly seen and a seventh segment which does not display such a prominent peak since it is amphiphilic. The hydrophobic regions depicted in the each correspond to the TM helices in bacteriorhodopsin for which there is a structure deposited on Brookhaven data-base. Bangham (1988) has developed a sieved version of the Kyte-Doolittle (KD) plot. While the KD method can be applied to detect transbilayer helices, Engelman *et al.* (1986) is credited with developing a method aimed specifically at identifying nonpolar transbilayer helices in amino acid sequences of proteins. These workers developed hydrophobicity scale calculated on the basis that the bilayer interior is a region of dielectric constant 2 containing no hydrogen bond donors or acceptors. It was also noted that details of helical structure must impact on the hydrophobicity scale computations.

4.3.3.7.2 Pattern-Matching Discriminators

Attwood *et al.* (1991) constructed a multiple alignment containing 37 sequences from related families of membrane bound receptors believed to share the same structural features as rhodopsin. Database pattern-scanning methods were then used to build a set of discriminators which can be used to identify each of the TM helices in GPCRs without regard to homology. The orientation of these helices in terms of exposure to the surrounding lipid or interiors was not considered.

4.3.3.7.3 Fourier Analysis of Multiple Sequence Format Files (MSF)

Komiya *et al.* (1988) and Rees *et al.* (1989b) worked on the structure determination of the photosynthetic reaction centre (RC) from *Rhodobacter sphaeroides* and *Rhodopseudomonas viridis*. These workers observed that the distribution of residues between different environments present within the membrane is non-random. Of the most abundant amino acids in the membrane, the apolar residues Leu, Ile, Phe and Val tend to be located on the side of the helix exposed to the membrane, whereas Trp, Thr, and Ser, show no particular preference between the interior and the surface faces of the TM helices of photosynthetic reaction centre. Comparison of aligned sequences from *Rhodobacter (Rb.) sphaeroides, Rb. capsulatus, Rhodopseudomonas viridis and R. rubrum* clearly showed that residues facing the interior were conserved in contrast to residues facing the lipid. Fourier transform methods (considered in more detail in chapter 8) were used to provide a quantitative approach for characterising the periodicity of conserved and variable residues in a family of aligned sequences. A periodicity of 3.4 residues per turn was observed in the MA helix (an ideal helix with residues conserved on just one face would have a periodicity of 3.6 residues per turn).

Using a sliding window of 19 residues in length and performing an averaged Fourier transform calculation, Rees *et al.* (1989b) correctly predicted the location of helices A and B of the reaction centre. Hence the Fourier transform method can be very successfully exploited in the form of a secondary structure prediction algorithm to locate TM helices which are amphipathic in character in terms of their residue conservation. The author has

coded up the several Fourier transform methods to provide useful tools in detecting and characterising amphipathic helices. A similar approach was very successfully adopted by Donnelly *et al.* (1989) as a suitable starting point for designing 3D models of GPCRs.

4.3.3.7.4 Helical and Beta Periodicity in Hydrophobicity

Eisenberg *et al.* (1984) noted that if the hydrophobic moment peak for a protein segment of amino acids peaks at or near 100^0 there is a likelihood of being an amphipathic helix. A segment with a hydrophobic moment that peaks at or near 180^0 has a likelihood of being a amphipathic β-segment. These implications have been applied in algorithms to identify secondary structure in the acetylcholine receptor (Finer-Moore and Stroud, 1984).

Cornette *et al.* (1987) have compared the discrete Fourier transform with a method based on least-squares fit of a harmonic sequence to a sequence of hydrophobicity values. 38 published hydrophobicity scales were examined for their ability to identify the characteristic period (3.6 residues per turn) of the α-helix. They found that the amphipathic index[7] is actually centred around 97.5^0 rather than the expected 100^0; though standard deviation of the location of the peaks for individual amphipathic helices is approximately 8^0. They concluded that the amphipathic index is a useful, objective measure of the ability of hydrophobicity scales to identify amphipathic helices.

4.3.3.7.5 Defining Helix Start and End Points

It is extremely difficult to decide the beginning and end positions of TM helices. In her assignment of GPCR helices, Joyce Baldwin (1993) noted that the best that could be hoped for was an accuracy of ±4 residues. Green (1990) in her review of structure prediction methods for membrane proteins reiterated that β-turns will tend to occur at the ends of TM helices. Hence the Chou and Fasman (1979) β-turn prediction method may be used to predict TM helix start and end points. However, a detailed analysis of this problem has led Donnelly and Cogdell (1993) to define a procedure for predicting the point at which

[7] The alpha amphipathic index measures the fraction of the total spectral area that is under the 97.5^0 peak.

a TM helix leaves the bilayer and penetrates the more polar region of the aqueous exterior. This was accomplished by comparing the relative directions of the hydrophobic and internal faces of the TM helices which should be contrasting only inside the bilayer. As they point out: "this information provides a strong constraint in the process of modelling membrane proteins". The Fourier transform method was used to monitor helical periodicity in hydrophobicity up to the point where the helix protrudes from the bilayer. By using a sliding window of either 10 or 12 residues to calculate the α periodicity index (*AP*) it is possible to predict the start and end points of the TM components of integral membrane proteins.

4.3.3.8 Energy Minimisation

The final phase of the combinatorial approach is the energy minimisation (EM) of the docked secondary structure components. Therefore, modelling packages which incorporate the ability to perform EM greatly aid the combinatorial approach[8]. Jameson (1989) points out that there are five dominant software packages at hand for molecular modelling: BioDesign Inc., Pasadena, California; Biosym Technologies, San Diego, California; Tripos, St. Louis, Missouri; Polygen, Waltham, Massachusetts; and Chemical Design Ltd., Oxford, UK. Typically, packages incorporate four features: (1) the ability to allow the user to quickly model complex structures, (2) a descriptive energy field, (3) an algorithm for performing molecular mechanics (i.e. EM) and (4) an algorithm for performing molecular dynamics (MD).

---

[8] The alternative approach is to save the generated 3D model and input it into a separate algorithm capable of performing the required EM. One such program being GROMOS (van Gunsteren, 1983). However, should any distances between non-bonded atoms approach ≤ 1 Å the EM will not converge and atoms will develop physically implausible vectors causing the structure to become unstable. QPACK (Gregoret and Cohen, 1990) is an algorithm which allows the user to identify bad contacts by slowly growing residues and observing which residues *touch*. Routine use of such algorithms just to identify bad contacts is likely to be tedious. Far more efficient to identify bad contacts quickly within the modelling package and then perform the EM, again within the modelling package. Modelling packages such as SYBYL incorporate colour force option which highlights atoms which are too close to one another allowing the user to adjust specific side-chain torsion angles (preferable to adjusting back-bone torsion angles) to remove any initial bad contacts thus allowing an EM run to converge rapidly on the first attempt. Yet another alternative is to use a function minimiser which smoothes out artificially large forces to avoid major structural perturbations allowing the user to then switch to a normal energy field description. Dauber-Osguthorpe *et al.* (1988) used a method of gradual annealing which involved applying a slight constraint to heavier atoms while allowing lighter atoms to relax.

The molecular modelling package from Tripos is known as SYBYL (Tripos E & S). It incorporates the Kollman united atom (Weiner *et al.*, 1984) and the Kollman all atom force field (Weiner *et al.*, 1986) which is especially designed to perform simulations of nucleic acids and proteins. (In the united atom force field the carbon hydrogens are collapsed inwards correspondingly increasing the carbon atoms van der Waal radia so that e.g. methyl groups are treated as single atoms thereby generating faster EM runs.) The Kollman force fields are also form an integral part of later versions of AMBER (Assisted Model Building with Energy Refinement - Weiner and Kollman, 1981) The force field equation for this force field is:

$$E_{\text{total}} = \sum_{\text{bonds}} K_R (R - R_\alpha)^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_\alpha)^2$$

$$+ \sum_{\text{dihedrals}} \frac{V_n}{2} [1 + \cos(n\phi - \gamma)]$$

$$+ \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}} \right]$$

$$+ \sum_{\text{H-bonds}} \left[\frac{C_{ij}}{R_{ij}^{12}} - \frac{D_{ij}}{R_{ij}^{10}} \right]$$

The first three terms represent the difference in energy between a geometry in which the bond lengths, bond angles and dihedral angles have ideal values and the actual geometry of the molecular structure. The fourth term represents the nonbonded van der Waals and electrostatic interactions. The final term (the so called 10-12 function) represents both strong and weak hydrogen bonding - also takes into account unrealistically short H-bonds. A full account of the coefficient values is given in the Weiner *et al.* (1984) paper. Of particular importance in the context of energy minimisations of the TM component of integral membrane proteins is the dielectric function ε. To partly take account of the

hydrophobic environment of the lipid bilayer a constant value in the range 2 to 5 is usually chosen in preference to a distance-dependent dielectric of say $\varepsilon = R_{ij}$.

Minimisation of the molecular structure not only requires an adequate field description but also a strategy needs to be employed to lower the value of the target function (for an excellent review of this subject see Mackay *et al.*, 1989). Three converging algorithms in popular use are: steepest descent, conjugate gradient and Newton-Raphson. A fourth less often used algorithm is simulated annealing. Space forbids detailed consideration of each method. In a nutshell, function minimizers consist of two major parts (a *heteroalgorithm*). There is a generic part: the *line search* (which actually changes the co-ordinates of the structure to a new lower-energy structure) and an algorithm (the different part, but there is *homology* between the gradient search methods) which decides the direction of the line search.

The distinguishing features of the different parts are:

- steepest descent - robust method but poor rate of convergence near the minimum. Users of this method notice that the energy of the structure is quick to fall, but algorithm is very poor at converging. This is because the each new direction vector can undo earlier progress at reaching convergence. For this reason, steepest descent is frequently followed by conjugate gradient to allow the energy of the structure to converge. Steepest descent relies on gradients to control the direction of the line search so that only 10-20% of function evaluations are required compared to more rigorous line search methods.

- conjugate gradients - minimises only along directions that are mutually conjugate. Returns a whole set of mutually conjugate gradient directions such that each successive step successively refines the direction toward the minimum. Frequently the method of choice for large systems.

- Newton-Raphson - makes use of the second derivative information (steepest descent and conjugate gradients methods only make use of first derivative information). Radius of

convergence and rate of convergence are poor in full blown Newton-Raphson. Block Diagonal Newton Raphson has a very large radius of convergence and has a superior rate of convergence compared to steepest descents algorithm.

- simulated annealing - minimisation by slow cooling. Structure is heated up to a high temperature and allow to cool slowly. Very good at avoiding multiple-minima problems or at least small barriers during the relaxation process and hence aid the search for the global mimima. Very similar in spirit to steepest descents except the energy is allowed to increase through fluctuations in the exchange of potential and kinetic energy.

4.4 Summary of Previous 3D Modelling Work

In critically reviewing papers devoted to modelling the DA family of GPCRs, two points should be firmly held in mind:

1. Humbler and Mizadegan (1992) in their review of 3D models of GPCRs point out that publications are few in number and lack detail - particularly x, y, z co-ordinates. Consequently, it is very difficult to compare the effectiveness of different modelling strategies.

2. This follows directly from the previous point. Given that 3D modelling studies devoted to the DA family of GPCRs is dearth of detail it is important to consider studies on closely related GPCRs particularly members of the β-adrenergic (since these are catechol amine binding GPCRs), muscarinic (binds the cationic amine acetylcholine) and serotonin families of GPCRs.

Joyce Baldwin reported to Richard Henderson (personal communication, 1994; appendix 2) that at a meeting in New Orleans: "she found there are literally hundreds of people making models of G-protein coupled receptors." This serves to underline the point that while there is a very strong interest in developing 3D models of GPCRs, there is an equal desire to avoid making the x, y, z co-ordinates available by depositing them on

Brookhaven[9]. This desire to just report in a fashion which lacks detail merely reflects the huge commercial interest in GPCRs in general and the cationic amine binding family of GPCRs in particular given their value as therapeutic targets for treating certain brain disorders.

4.4.1 Dahl *et al,* 1991a and 1993

In their model of the D_2 receptor, Dahl *et al.* (1991a) designed his model on the basis of five hypotheses: (i) TM segments are α-helices; (ii) each TM helix is 27 amino acids in length; (iii) TM helices can be predicted on the basis of multiple sequence alignments - i.e. homology is high in the putative TM regions; (iv) the putative α-helices are orientated with their polar surface areas facing inwards into the central core of the receptor; (v) GPCRs have a common ligand binding site. Dahl made much use of site directed mutagenisis studies to guide the docking of the DA agonist. Dahl's attempt at modelling the 3D structure of the dopamine receptor falls into the combinatorial approach to protein modelling in that he did not succumb to the temptation of using bacteriorhodopsin as a structural template. A similar approach was used to model the 5-HT_{1a} receptor (Sylte, *et al.,* 1993).

4.4.2 Maloney-Huss and Lybrand (1992)

Maloney-Huss and Lybrand (1992) in their modelling studies of the β_2 adrenergic receptor (which is closely related to all of the DA receptors) sought to answer to two important questions. Firstly, is it possible to construct a physically credible 3D model using the experimental data as a guide. If so, how could these models assist in furthering the study of these receptors. In their modelling efforts they were particularly mindful of the fact that previous modelling work by others had rarely sought to check the plausibility of the constructed models against available experimental data. Also, these workers set out to model the entire protein, something which others considered impossible. For example, in

[9] Brookhaven accepts 3D models; for example: a 3D model of a four helix-bundle called Felix (Hecht *et al.,* 1990) has been deposited on Brookhaven and is freely available to the biotechnology community.

their 3D modelling of the M_1 receptor Saunders and Findlay (1990) completely ignored the 150-250 residue third intracellular loop connecting helices V and VI: "because we do not know how to model this region".

Secondary prediction algorithms such as the Kyte and Doolittle (1982) hydropathy analysis method together with sequence alignments with other GPCRs was used to predict secondary structure. Overall topology chosen was identical to that proposed by Engelman *et al.* (1980) and Findlay (Ryba *et al.*, 1992), i.e. TM helices were judged to be located in anti-parallel fashion in numerical order. This followed from antibody mapping experiments (Wang *et al.*, 1989) that localised the carboxy-tail and interhelical loop segments between helices 1-2, 3-4, 5-6 on the intracellular side of the membrane and loop segments between helices 2-3, 4-5, 6-7 (plus amino terminus) on the extracellular side. Also, amphiphilic helices were predicted at the both ends of iIII.

The individual TM helices were modelled, but no details were given concerning the modelling of kinked helices (in particular helices II, V and VI had middle Pro residues). The lack of sequence homology between the proposed heptahelical motif in β_2 and the seven helix bundle observed in the intermediate structure of bacteriorhodopsin suggested that bacteriorhodopsin should not be used as a structural template. Hydrophobic moments were used to guide the orientation of each helix relative to the surrounding lipid (i.e. helix phase). Energy minimisation *in vacuo* was used in a limited form to avoid compaction problems to improve helix packing. Loops were initially modelled by manually growing a loop from the end of one helix by setting all ϕ/ψ angles set to 180^0. Then, using 2 to 5 ps of a novel low temperature molecular dynamics (MD) generated allowable ϕ/ψ and χ torsion angles. A weak harmonic constraint was used to pull the free end of the loop fragment into place to form a *trans* peptide bond with the amino end of the second target helix. A similar constraint was presumably applied to allow a disulphide bond to form between the conserved Cys residues of iI and iII. The amino and carboxy terminal fragments were modelled using the work of Ponder and Richards (1987) to establish reasonable χ torsion angles.

Maloney-Huss and Lybrand considered their modelling fell into the *de novo* category of protein modelling since they did not succumb to the temptation to using bacteriorhodopsin as a structural template to locate the TM helices of their model of β_2. However, it could be argued that their modelling style essentially followed the combinatorial method developed by Fred Cohen (Cohen, 1979).

4.4.3 Hibert *et al.* (1992)

This group succumbed to the temptation to use bacteriorhodopsin as a structural template despite the lack of sequence homology with GPCRs. Helices were individually constructed using the $\phi\psi$ angles (-59^0 and -44^0) recommended for helical residues in a non-polar environment (Blundell *et al.*, 1983). Details concerning the modelling of kinks due to conserved Pro residues were not described. Helix phase was judged by orientating conserved residues towards the centre of the protein. Extensive energy minimisation was applied with no concern expressed about possible compaction problems or whether the energy of the models converged. Given that extensive energy minimisation was used it would have been wise to have made use of QPACK (Gregoret and Cohen, 1990), an algorithm freely available to evaluate protein structures in terms of packing densities. In this manner Hibert's group (Trumpp-Kallmeyer *et al.*, 1992) set about modelling the TM regions of 39 GPCRs including the DA receptor family. Hibert used his models to announce (despite considerable experimental evidence to the contrary) that the conserved Asp residue (conserved in the catechol amine binding GPCRs only) in the middle of TM helix II only plays a structural and allosteric functional role. This conflicts completely with the views of Findlay group at Leeds (for example: Saunders and Findlay, 1990) who suggest that this Asp residue plays a vital role in the binding of the agonist. Henderson (appendix 2) states quite clearly that the use of the bacteriorhodopsin to model GPCRs is flawed since the helices are up to 10Å apart (Gebhard *et al.*, 1993). Indeed, Hibert has been obliged to counter strong criticism of his modelling strategy in the literature (Hibert *et al.*, 1993). Since Hibert has used the bacteriorhodopsin as a structural template it naturally follows that he has adhered to the homology modelling approach to model the DA family of receptors despite complete lack of any sequence homology between bacteriorhodopsin and the DA family of receptors.

4.4.4 Cronet *et al.* (1993)

Their modelling procedure involved three levels of detail: overall topography, location of helix ends and helix phase. Hydropathy analysis was used to predict helix end positions to ±4 residues accuracy and by considering experimental evidence relating to the β_2-adrenergic receptor. The rotational orientation of the helices was determined using environmental preference parameters which were derived from bRh and three homologues. The parameters were then used to determine the optimal fit of the β_2-adrenergic receptor onto the bRh structural template. Energy minimisation was used to optimise the TM packing. Hence, these workers also succumbed to the temptation to use bacteriorhodopsin as a structural template. Consequently, their modelling strategy is now obsolete bearing in mind the recent work of Baldwin (1993) on the probable arrangement of the helices in G protein-coupled receptors based on the projection structure of rhodopsin which is significantly different from that of bacteriorhodopsin (Gebhard *et al.*, 1993; also see section 3.2 and figure 3.2.1). However, Cronet *et al.* (1993) make it clear that their model is available on request and so confounds the scepticism of Humbler and Mizadegan (1992).

5. Sequence Analysis of Catechol Amine Binding G Protein Coupled Receptors

5.1 Summary

Extensive homology analysis using the Needleman and Wunsch (1970) algorithm failed to detect any significant sequence homology between bacteriorhodopsin (bR) and any catechol amine binding G protein coupled receptor (GPCR). However, extensive hydropathy analysis together with the distribution of potential tripeptide glycosylation sequences suggests that the overall topology of these GPRCs and bR is similar. That is, catechol amine binding GPCRs appear to have a heptahelical transmembrane motif similar to that of bR. Members of the D_1 subfamily of dopamine GPCRs appear to have a short third intracellular loop and long carboxy tail. Conversely, members of the D_2 subfamily of dopamine receptors appear to have a long third intracellular loop and a short carboxy tail. Secondary structure prediction algorithms failed to identify transmembrane helices correctly. However, the profile network prediction algorithm (PHD; Rost *et al.*, 1994) predicted a short α-helix both ends of the third intracellular loop in each member of the dopamine (DA) family of GPCRs. PHD predicts the short carboxy tail of the D_2 sub-family of DA receptors to be a α-helix. Also, a short α-helix (three turns) between the carboxy-terminal end of putative transmembrane helix VII and the first Cys residue in the long carboxy tail of the D_1 sub-family (D_1 and D_5).

5.2 Introduction

Only recently have the primary structures of the G protein coupled receptors (GPCRs) been illucidated. While the number of GPCRs sequenced now runs into hundreds a three-dimensional model does not exist of any GPCR. The business of obtaining crystals of native membrane proteins capable of diffracting X-rays encompasses considerable problems. In the absence of a 3D structure the first step in modelling the dopamine (DA) family of receptors (in particular D_2) must begin with a careful analysis of the primary

structure - the sequence of amino-acids that makes up the DA G protein coupled receptor (GPCR).

The sponsor, Organon (Scotland) Ltd. was particularly interested in using bacteriorhodopsin (bRh), a known structure (Henderson and Unwin, 1975; Henderson *et al.*, 1990), as a template to model D_2. bRh is an integral membrane protein with a hepta-helical transmembrane region. The transmembrane helices of bRh are arranged in an anti-parallel fashion with the amino-tail exterior to the cell membrane and the short carboxy-tail interior to the membrane. Given that the sponsor expressed a strong desire for homology modelling of the D_2 receptor using bRh as a template it is sensible to assess the scientific validity of this intended procedure.

While the transmembrane region of bRh and D_2 might indeed share resemblances. The loop regions of the DA family of G protein coupled receptors (GPCRs) would certainly bear no resemblance to the loop regions of bRh. The explaination for this simply being that the loop regions of each DA GPCR are considerably longer than the loop regions of bRh as indicated by simple hydropathy analysis. Hence the use of accurate secondary structure prediction tools such as PHD (Profile Network Prediction Heidelberg; (Rost and Sander, 1993; Rost *et al.*, 1994) makes good sense. PHD claims 80% accuracy when applied to non-membrane proteins. Given that the loop DA GPCRs are in the extracellular or intracellular environment, the rules of folding of these regions are likely to correspond closely to the rules governing the folding of globular proteins.

5.3 Methods and Materials

5.3.1 Homology Analysis

Homology analysis was carried out using the GAP program (Devereux, *et al.*, 1984) and was applied to a range of GPCRs (see table 5.2.1) and bRh. Default values were used for gap weight and gap length weight: 3.00 and 0.1 in units of matched bases. [Gap uses the alignment method of Needleman and Wunsch, 1970]. The RAN option of GAP

was used to calculate the average alignment score and standard deviation from 100 randomized alignments in which the second sequence was repeatedly shuffled while maintaining the length and composition of the original sequence. In this way Z scores were calculated (equation 1).

$$Z = (\chi - m) / \sigma \qquad \text{Eq (1)}$$

where χ is the local alignment score, m is the mean of random scores and σ is the standard deviation of random scores (Dayhoff, 1978; Doolittle, 1981; Lipman and Pearson, 1985). Z scores greater than 5 should be regarded as significant (Bell, 1992).

GPCR	FETCH NO.	Abbrev	Key Reference
Rhodopsin	P08100	rH	Nathans and Hogness, 1984
D_1	P02945	D1	Dearry *et al.*, 1990
D_2	P21728	D2	Dal-Toso *et al.*, 1989
D_3	P14416	D3	Sokoloff *et al.*, 1990
D_4	P19020	D4	Van-Tol *et al.*, 1991
D_5	P21917	D5	Sunahara *et al.*, 1991
α_1 adrenergic	P21918	A1	Libert *et al.*, 1989
α_{A2A} adrenergic	P08913	A2A	Fraser *et al.*, 1989
α_{A2B} adrenergic	P18825	A2B	Regan *et al.*, 1988
α_{A2C} adrenergic	P18089	A2C	Lomasney *et al.*, 1990
β_1 adrenergic	P08588	B1	Frielle *et al.*, 1987
β_2 adrenergic	P07550	B2	Schofield, *et al.*, 1987
β_3 adrenergic	P13945	B3	Emorine *et al.*, 1989
5-HT-1A	P08908	5HT	Kobilka, *et al.*, 1987
M_1	P11229	M1	Allard *et al.*, 1987

Table 5.2.1 GPCR primary structures used in homology analysis. First column lists GPCR type; second column the SWISSPROT FETCH Number; the third column the abbreviation used in the remainder of this chapter; fourth column lists ealiest key reference. All of the primary structures were human, except for: D_3 (rat) and α_1 adrenergic (dog). The remaining primary structure used was: bacteriorhodopsin (P02945); code: bRh (Dunn *et al.*, 1981).

5.3.2 PEPPLOT Analysis

The PEPPLOT command (GCG sequence analysis software - Devereux, *et al.*, 1984) was applied to the bacteriorhodopsin sequence (bRh) and D_1 and D_2 sub-type receptors. Hydropathy analysis of primary structures was performed on each member of the DA family of GPCRs using the Kyte-Doolittle or KD plot (Kyte and Doolittle, 1982) and the Goldman, Engelman and Steize or GES curve (often quoted as: Goldman *et al.*, 1986; properly quoted as: Engelman *et al.*, 1986). The KD plot used the default window size of nine residues. The secondary structure prediction method of Chou and Fasman (1978) was used to calculate the propensity measures for alpha-helix and beta-sheet. Both curves are the average of a residue-specific attribute over a window of four. Turns were predicted using the method of Chou and Fasman (1978). The curve is the product of a residue-specific, position-dependent attribute (probability) multiplied across a window of four. The calculated values are multiplied in Pepplot by 10,000 for plotting.

5.3.3 Tripeptide Glycosylation Sequences in Dopamine GPCRs

Each primary structure was examined for the occurrence of the tripeptide sequence -Asn-X-Ser or -Asn-X-Thr (i.e. -Asn-X-Ser/Thr), where X represents one of the 20 amino acids (Eylar, 1965; Marshall, 1974).

5.3.4 Profile Network Prediction Heidelberg - PHD

PHD (Rost and Sander, 1993; Rost *et al.*, 1994) was applied to each member of the dopamine family of GPCRs to predict α-helix (H), β-strand (E) and loop (L). PHD was accessed by emailing primary structures to: PredictProtein@EMBL-Heidelberg.DE. Secondary structure predictions for each of the dopamine receptors was automatically returned in about 4 hours.

5.4 Results

GAP results are displayed in tables 5.3.1 to 5.3.5 (pages 81-85). PEPPLOT results are diplayed in figures: 5.3.1 to 5.3.4 (pages 86-89). The location of potential glycosylation sites in the dopamine family of GPCRs is given in table 5.3.6 (page 90). PHD results are displayed in figures 5.3.5 to 5.3.9 (pages 91-100).

	bRh	rH	D1	D2	D3	D4	D5	A1	A2A	A2B	A2C	B1	B2	B3	5HT	M1
bRh	100	20	20	18	16	18	16	16	18	19	18	22	19	20	17	18
rH	22	100	23	27	30	22	25	20	22	22	22	22	21	24	24	25
D1	20	23	100	42	37	34	59	33	38	36	37	35	34	34	41	30
D2	18	27	42	100	54	43	40	41	36	33	31	36	38	34	34	26
D3	16	30	37	54	100	42	33	39	36	33	32	37	34	37	35	30
D4	18	22	34	43	42	100	38	34	37	41	38	38	30	40	33	32
D5	16	25	59	40	33	38	100	33	38	39	40	38	31	34	43	29
A1	16	20	33	41	39	34	33	100	37	38	39	36	33	36	43	32
A2A	18	22	38	36	36	37	38	37	100	55	54	39	34	35	34	29
A2B	19	22	36	33	33	42	39	38	55	100	55	36	33	35	34	29
A2C	18	22	37	31	32	38	40	39	54	55	100	35	32	36	33	28
B1	22	22	35	36	37	38	38	36	39	36	35	100	52	55	39	30
B2	19	21	34	38	34	30	31	33	34	33	32	52	100	47	34	31
B3	20	24	34	34	37	40	34	36	35	35	36	55	47	100	36	32
5HT	17	24	41	34	35	33	43	43	34	34	33	39	34	36	100	30
M1	18	25	30	26	30	32	29	32	29	29	28	30	31	32	30	100

Table 5.3.1 Percentage homology matrix; it is clear that the percentage homology between bRh (bacteriorhodopsin) and members of the dopamine, alpha and beta adrenergic families, serotonin ($5HT_{1A}$), and muscarinic M1 G protein coupled receptors (GPCRs) is too low to justify homology modelling using bRh as a template to model the dopamine GPCRs.

	bRh	rH	D1	D2	D3	D4	D5	A1	A2A	A2B	A2C	B1	B2	B3	5HT	M1
bRh	77	2.2	0.9	0.2	0.5	1.5	0.5	0.3	1.1	0.3	1.9	0.2	0.5	1.3	2.0	3.7
rH	1.9	106	8.7	8.9	8.0	8.7	8.4	1.3	5.7	5.3	2.6	9.2	7.7	12	8.3	6.6
D1	1.5	7.6	105	20	16	18	61	16	23	16	23	27	27	22	23	14
D2	0.0	8.6	18	117	44	39	15	14	27	29	26	22	16	18	26	20
D3	0.8	8.3	15	50	138	30	13	13	31	25	24	15	18	19	25	17
D4	1.2	8.2	16	35	37	108	19	9.7	25	27	22	21	16	19	23	20
D5	0.4	9.0	68	16	14	21	120	16	18	16	14	29	27	22	22	13
A1	0.5	0.9	17	14	11	8.1	15	109	12	14	13	20	18	13	15	8.4
A2A	0.1	6.6	16	32	28	26	17	12	97	49	45	18	17	16	26	19
A2B	0.6	5.5	13	33	25	25	17	12	44	90	43	19	18	19	32	20
A2C	0.8	2.9	15	28	22	24	15	12	49	56	107	17	13	15	23	18
B1	0.2	11	24	19	16	22	25	16	19	19	19	118	52	38	24	18
B2	0.3	7.9	29	18	16	16	31	17	17	19	19	48	104	38	23	15
B3	1.8	12	22	19	18	23	26	14	21	21	17	47	42	106	25	18
5HT	1.7	6.8	21	24	22	22	21	13	29	32	21	23	19	24	102	22
M1	4.2	5.3	12	20	17	16	12	7.5	22	25	20	18	15	15	21	110

Table 5.3.2 Z scores matrix. Any value above 5 indicates significant sequence homology (Bell, 1992). Values less than 10 are given to one decimal place. It is clear that bacteriorhodopsin (bRh) has no significant sequence homology with any member of the dopamine, alpha and beta-adrenergic families (or serotonin $5HT_{1A}$ or muscarinic M1 receptor). However, rhodopsin (rH) has significant sequence homology with each member of the dopamine family of receptors as well as other G protein coupled receptors (GPCRs). Hence, the case for using bRh in homology modelling of dopamine (DA) GPCRs is not justified.

	bRh	rH	D1	D2	D3	D4	D5	A1	A2A	A2B	A2C	B1	B2	B3	5HT	M1
bRh	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rH	0	1	1	1	1	1	1	0	1	1	0	1	1	1	1	1
D1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
D2	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
D3	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
D4	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
D5	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
A1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
A2A	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
A2B	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
A2C	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
B1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
B2	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
B3	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5HT	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
M1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 5.3.3 Binary verion of table 5.3.2; binary 1 signifies Z scores greater than 5. It is clear that bRh has no significant homology with any of the GPCRs used in this homology study. rH has significant homology with every member of the DA family of GPCRs. Of the GPCRs examined, rH only lacks significant homology with the alpha-adrenergic receptors: A1 and A2C.

	bRh	rH	D1	D2	D3	D4	D5	A1	A2A	A2B	A2C	B1	B2	B3	5HT	M1
bRh	1															
rH		1														
D1			1	1			1		1		1	1	1	1	1	
D2				1	1	1			1	1	1	1			1	1
D3				1	1	1			1	1	1				1	
D4				1	1	1			1	1	1	1			1	1
D5			1			1	1					1	1	1	1	
A1								1				1	1			
A2A				1	1	1			1	1	1				1	
A2B				1	1	1			1	1	1				1	1
A2C				1	1	1			1	1	1				1	
B1			1			1	1					1	1	1	1	
B2			1				1					1	1	1	1	
B3			1			1	1		1	1		1	1	1	1	
5HT			1	1	1	1	1		1	1	1	1		1	1	1
M1				1					1	1	1				1	1

Table 5.3.4 Binary version of table 5.3.2 accept that boxes with zeros are just left blank. Binary 1 signifies Z scores greater than 20. Using this cut-off value reveals a definite clustering effect. For example, the D_2 sub-family of DA receptors have particularly high homology with the alpha-adrenergic family of GPCRs. This rather suggests that three dimensional models of any of D_2, D_3 or D_4 would be useful starting points to model members of the alpha-adrenergic family of GPCRs and visa-verca.

	bRh	rH	D1	D2	D3	D4	D5	A1	A2A	A2B	A2C	B1	B2	B3	5HT	M1
bRh	1															
rH		1														
D1			1				1									
D2				1	1	1										
D3				1	1	1			1							
D4				1	1	1										
D5			1				1									
A1								1								
A2A				1					1	1	1					
A2B				1					1	1	1				1	
A2C									1	1	1					
B1												1	1	1		
B2							1					1	1	1		
B3												1	1	1		
5HT										1					1	
M1																1

Table 5.3.5 Binary version of table 5.3.2; binary 1 signifies Z scores greater than 30. Using this cut-off value produces a definite clustering effect revealing individual families of GPCRs. Also, the two sub-families in the DA family of GPCRs are clearly differentiated: the D_2 sub-family: D_2, D_3 and D_4 ; and the D_1 sub-family: D_1 and D_5. It is also clear for example that a three dimensional model of 5HT (5-HT-1A) could be used as a starting template for modelling A2B (and visa-versa) since 5HT has particularly high homology with A2A.

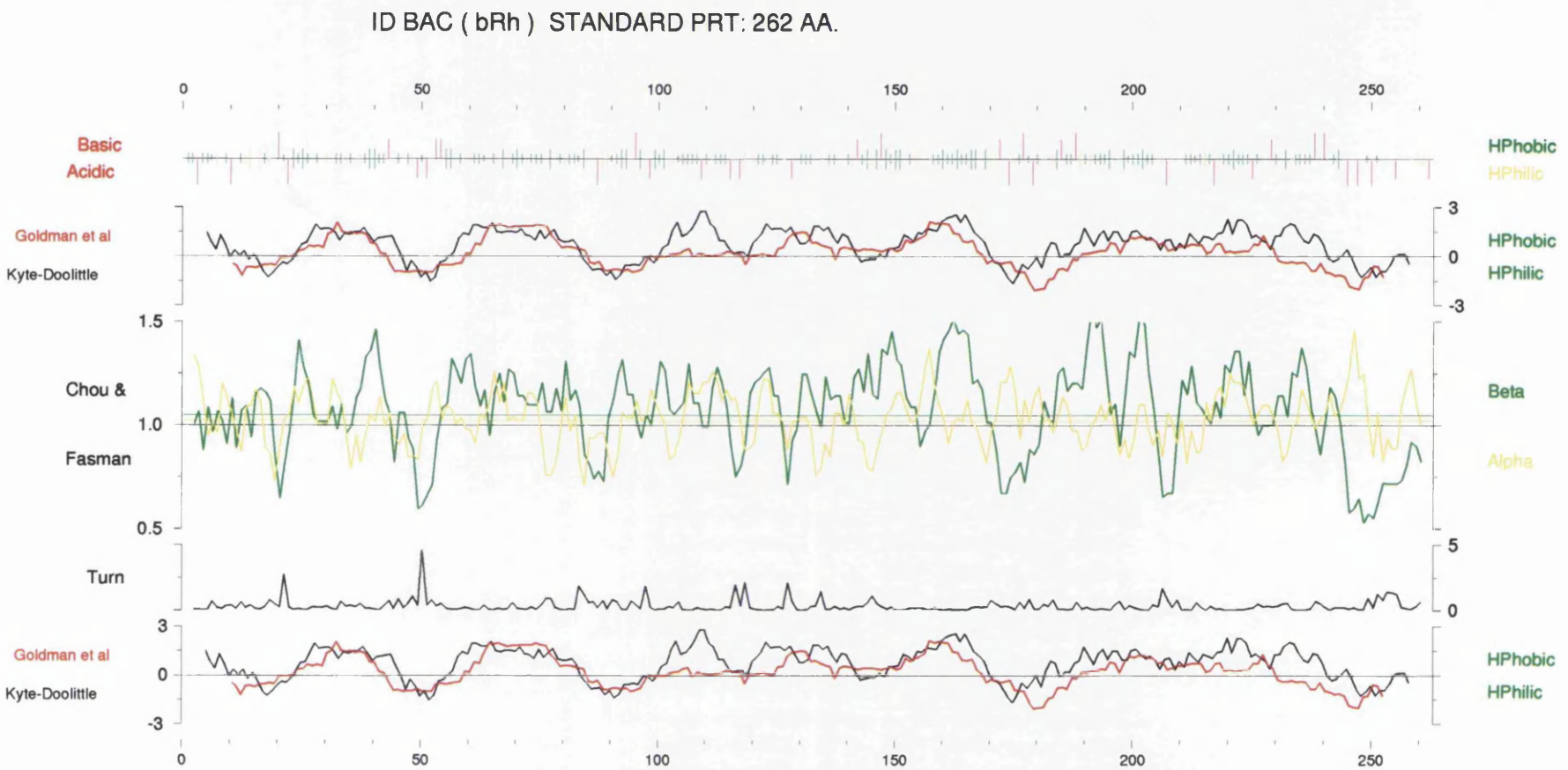


Figure 5.3.1 Application of PEPPLOT (GCG sequence analysis software - Devereux, *et al.*, 1984) to the bacteriorhodopsin sequence (bRh). The first (i.e. top panel) panel shows the location of acidic, basic, hydrophilic and hydrophobic residues. Panel two is a sliding window hydropathy analysis of bRh using the Kyte-Doolittle or KD plot (Kyte and Doolittle, 1982) and the Goldman, Engelman and Steize or GES curve (often quoted as: Goldman *et al.*, 1986; properly quoted as: Engelman *et al.*, 1986). The green curve in the lowest panel is the GES curve for identifying nonpolar transbilayer helices. The third panel shows the Chou and Fasman (1978) propensity measures for α-helix and β-sheet. The fifth panel is merely a repeat of panel two to aid the reader.

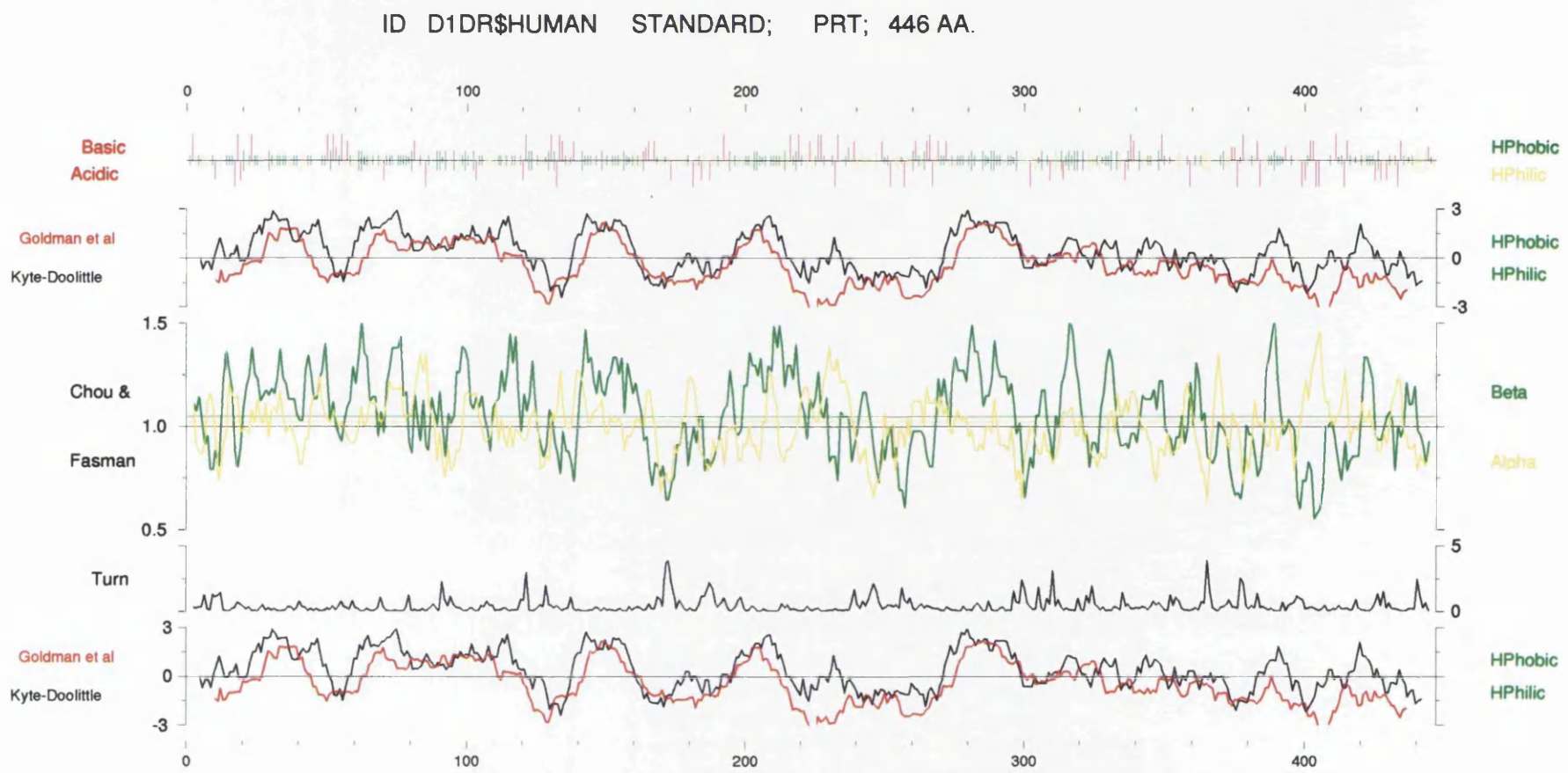


Figure 5.3.2 Application of PEPPLOT (GCG sequence analysis software - Devereux, *et al.*, 1984) to the human D_1 DA receptor. The legend is the same as that for figure 5.3.1. D_1 (like D_5) has a short third intracellular loop between transmembrane helices V and VI. This is reflected in the hydropathy analysis where there is a gap between the prominent peaks coresponding to helices V and VI. Transmembrane helix does not show up clearly in the hydropathy analysis indicating that it is more hydrophilic than the other transmembrane helices. Also, the D_1 subfamily of DA receptors is characterised by a long intracellular carboxy tail.

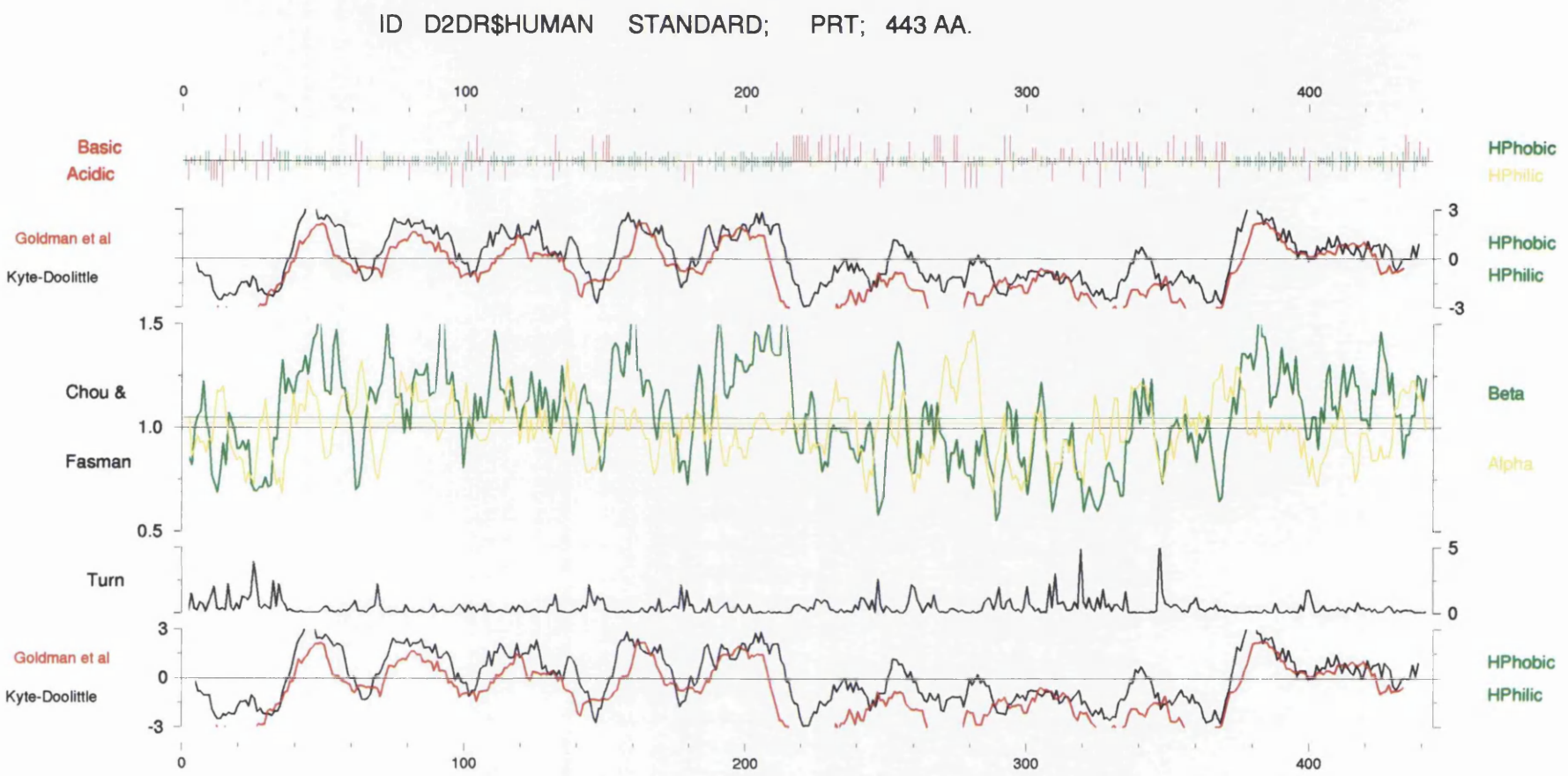


Figure 5.3.3 Application of PEPPLOT (GCG sequence analysis software - Devereux, *et al.*, 1984) to the human D_2 DA receptor. The legend is the same as that for figure 5.3.2 except that D_2 (like D_3 and D_4) has a long third intracellular loop between transmembrane helices V (TM5) and VI (TM6) and a shorter tail.

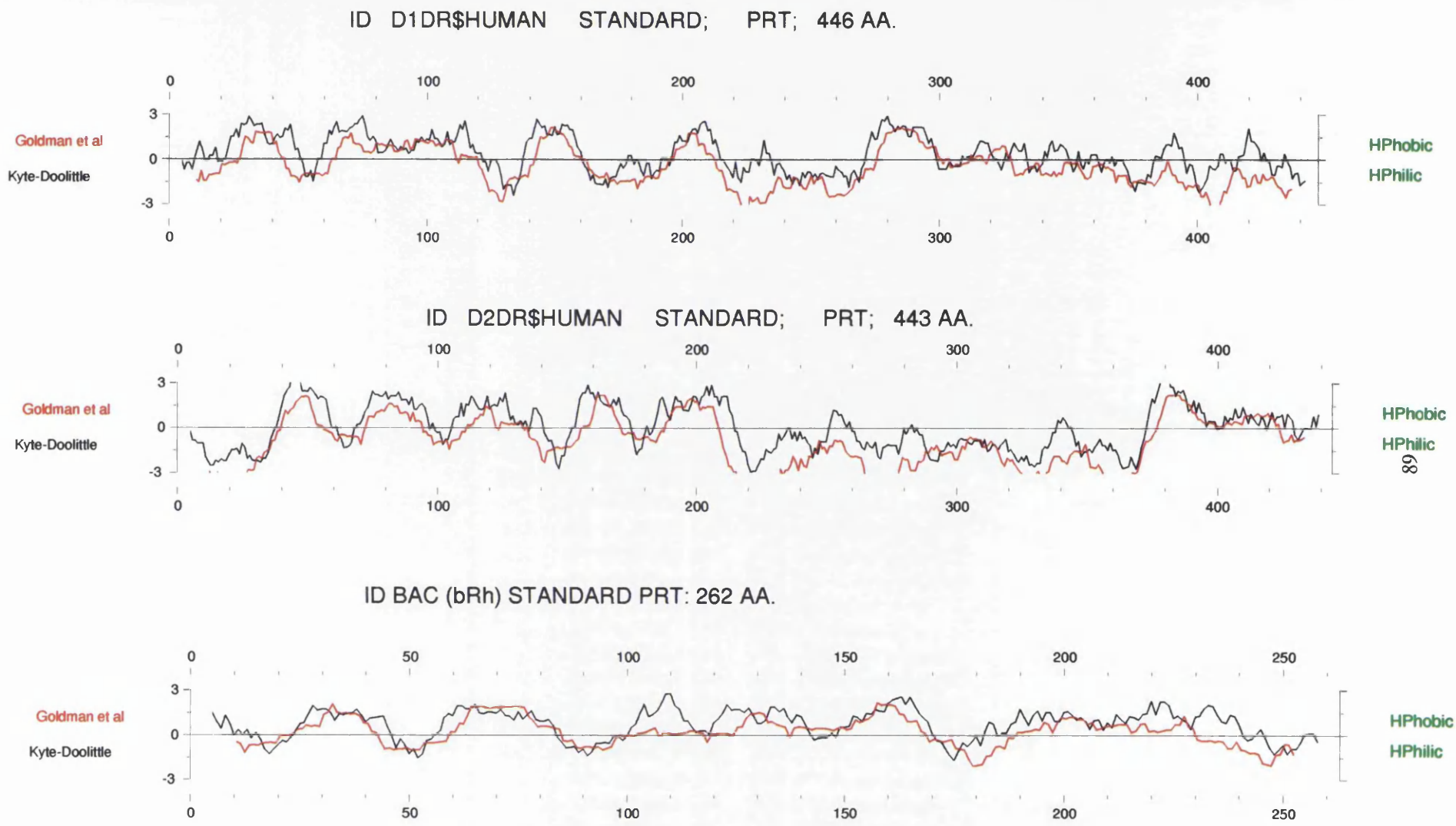


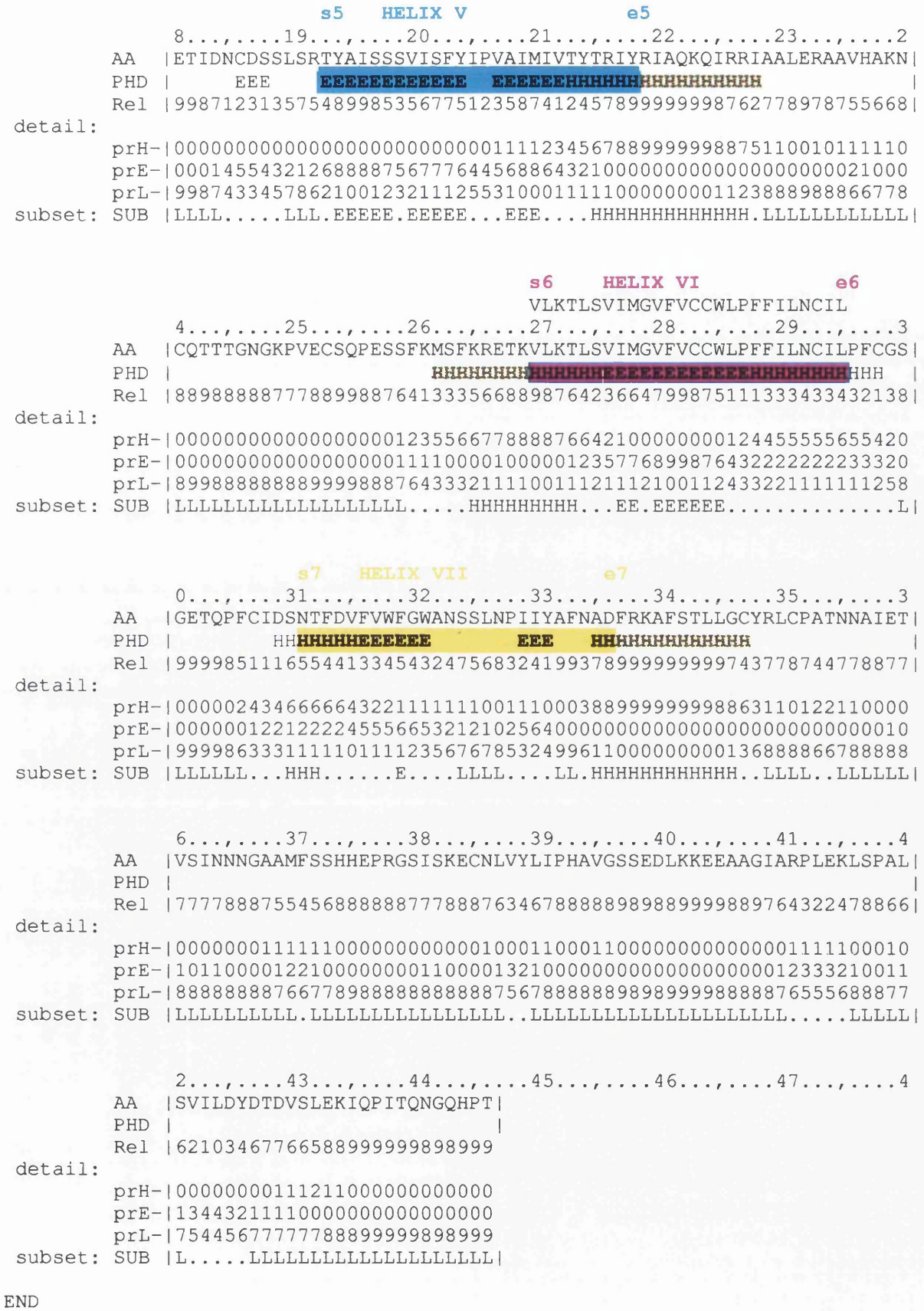
Figure 5.3.4 Comparison of KD and GES plots of D_1 and D_2 sub-type DA receptors and bRh. It is clear that the first six transmembrane helices of the D_1 and D_2 sub-type receptors are reasonably easy to identify, but transmembrane helix VII is clearly less hydrophobic. The much longer stretch of more hydrophilic residues between the fifth and sixth peaks (coresponding to the third intracellular loop) in D_1 is also clearly discernable. Likewise D_1 sub-type has a much longer tail sequence than that found in the D_2 subtype DA receptor. A long intracellular loop and carboxy tail is clearly not found in the bRh.

Receptor	Position (of Asn residue)	Tripeptide Sequence (Single letter code)	Location	Extracellular?
D1	5	**NTS**	AMINO TAIL	YES
D2	5	**NLS**	AMINO TAIL	YES
D2	17	**NWS**	AMINO TAIL	YES
D2	23	**NGS**	AMINO TAIL	YES
D3	12	**NST**	AMINO TAIL	YES
D3	19	**NST**	AMINO TAIL	YES
D3	173	**NTT**	eII	YES
D4	3	**NRS**	AMINO TAIL	YES
D5	7	**NGT**	AMINO TAIL	YES

Table 5.3.6 Location of potential glycosylation sites in the dopamine family of GPCRs. Tripeptide sequences of interest are: -Asn-X-Ser or -Asn-X-Thr (i.e. -Asn-X-Ser/Thr), where X represents one of the 20 amino acid. Whether a site is glycosylated depends on where this tripeptide sequence is accessible ot the oligosachharide transferase (e.g. on exposed coils). The amino-tail is in the extracellular medium and likewise for eII (the second extracellular loop) - see figure 3.2 (page 27).

Figure 5.3.5 Profile Network Prediction Heidelberg - PHD; a secondary prediction tool - here applied to the primary structure of the D_1 receptor. Abbreviations: secondary structure: H=helix, E=extended (sheet), blank=rest (loop); AA: amino acid sequence; PHD: Profile network prediction HeiDelberg; Rel: Reliability index of prediction (0-9); detail: prH: 'probability' for assigning helix, prE: 'probability' for assigning strand, prL: 'probability' for assigning loop note: the 'probabilites' are scaled to the interval 0-9, i.e. prH=5 means, that the signal at the first output node is 0.5-0.6; subset: SUB: a subset of the prediction, for all residues with an expected accuracy > 82% note: for this subset the following symbols are used: L: is loop (for which above " " is used), ".": means that no prediction is made for this residue, as the reliability is Rel < 5; s1: start of helix I, e1: end of helix 1. Putative transmembrane helices corespond to those suggested by Baldwin (1993). It is clear that PHD predicts mainly β-sheet for each transmembrane segment. However, PHD predicts α-helix secondary structure at the each end of the third intracellular loop (iIII). Also, a short α-helix (three turns) between the carboxy-terminal end of putative transmembrane helix VII and the first Cys residue in the long carboxy tail.

```
# D1_human          length      446 amino acids

                                      s1       HELIX I         e1       s2 H2
              ....,....1....,....2....,....3....,....4....,....5....,....6
        AA   |MRTLNTSAMDGTGLVVERDFSVRILTACFLSLLILSTLLGNTLVCAAVIRFRHLRSKVTN|
        PHD  |                          E                          EE      E|
        Rel  |997688888988876778874058887777889999984148999873222221689743|
detail:
        prH-|00110100000000000000000001110000000000000000001111244200001 1
        prE-|00000000000000111001346888777788899988646899987555422310012 5
        prL-|99888888998888788888642000011110000001353000001223222378875 3
subset: SUB |LLLLLLLLLLLLLLLLLLL..EEEEEEEEEEEEEEE...EEEEEE.......LLLL..|

                 HELIX II        e2              s3        HELIX III
              ....,....7....,....8....,....9....,....10...,....11...,....1
        AA   |FFVISLAVSDLLVAVLVMPWKAVAEIAGFWPFGSFCNIWVAFDIMCSTASILNLCVISVD|
        PHD  |EEEEEEEE  EEEEEEEE  EEEEEEEEEE   EEEEEEEEEEEEEE EEEEEEEE |
        Rel  |787654321115778986231322454555479647889963334432202455453 3|
detail:
        prH-|10111222322100000001222332121110000000000001111232421122222 2
        prE-|78766554334678888732344455656666102688899755555544345666654 2
        prL-|10011223333110000255322111111212797210000233222222222111123 4
subset: SUB |EEEEE......EEEEEEE........E.EEE.LLL.EEEEEE..........EE.E...|

                e3              s4    HELIX IV          e4
              2...,....13...,....14...,....15...,....16...,....17...,....1
        AA   |RYWAISSPFRYERKMTPKAAFILISVAWTLSVLISFIPVQLSWHKAKPTSPSDGNATSLA|
        PHD  |HHHHHH HHH       EEEEEEEEEEEEEEEE                         |
        Rel  |144431133214565722125789999999999634641134788999998877756457|
detail:
        prH-|45555445543222213442111000000000000023322000000000000011221 1
        prE-|22221110011100101124678888989999876311222110000000000001000 0
        prL-|32122343334677675332110000000000236764346788999998888877678 8
subset: SUB |............LLLL....EEEEEEEEEEEEE..L.....LLLLLLLLLLLLLL.LL|
```


s5 HELIX V e5
8...,....19...,....20...,....21...,....22...,....23...,....2
AA |ETIDNCDSSLSRTYAISSSVISFYIPVAIMIVTYTRIYRIAQKQIRRIAALERAAVHAKN|
PHD | EEE EEEEEEEEEEEE EEEEEEHHHHHHHHHHHHHHH |
Rel |998712313575489985356775123587412457899999998762778978755668|
detail:
prH-|000000000000000000000000011112345678899999988751100101111110
prE-|000145543212688887567776445686432100000000000000000000021000
prL-|998743345786210012321112553100011111000000001123888988866778
subset: SUB |LLLL.....LLL.EEEEE.EEEEE...EEE....HHHHHHHHHHHHH.LLLLLLLLLLL|
s6 HELIX VI e6
VLKTLSVIMGVFVCCWLPFFILNCIL
4...,....25...,....26...,....27...,....28...,....29...,....3
AA |CQTTTGNGKPVECSQPESSFKMSFKRETKVLKTLSVIMGVFVCCWLPFFILNCILPFCGS|
PHD | HHHHHHHHHHHHHHHHEEEEEEEEEEEEEHHHHHHHHHHHHH |
Rel |889888887778899887641333566898764236647998751113334334321381
detail:
prH-|000000000000000001235566778888766421000000001244555556554200
prE-|000000000000000001111000010000012357768998764322222222333200
prL-|899888888889999888764333211110011121112100112433221111111258
subset: SUB |LLLLLLLLLLLLLLLLLL.....HHHHHHHHH...EE.EEEEEE.............L|
s7 HELIX VII e7
0...,....31...,....32...,....33...,....34...,....35...,....3
AA |GETQPFCIDSNTFDVFVWFGWANSSLNPIIYAFNADFRKAFSTLLGCYRLCPATNNAIET|
PHD | HHHHHHHEEEEEE EEE HHHHHHHHHHHHHH |
Rel |999985111655441334543247568324199378999999999743778744778877|
detail:
prH-|000002434666664322111111100111000388999999998863110122110000
prE-|000000122122224555665321210256400000000000000000000000000010
prL-|999986333111110111123567678532499611000000001368888667888888
subset: SUB |LLLLLL...HHH......E....LLLL....LL.HHHHHHHHHHHH..LLLL..LLLLLL|
6...,....37...,....38...,....39...,....40...,....41...,....4
AA |VSINNNGAAMFSSHHEPRGSISKECNLVYLIPHAVGSSEDLKKEEAAGIARPLEKLSPAL|
PHD | |
Rel |777788875545688888877788876346788888989889999889764322478866|
detail:
prH-|000000011111100000000000010001100011000000000000001111100010
prE-|101100001221000000001100001321000000000000000000012333210011
prL-|888888887667789888888888887567888888989899999888876555688877
subset: SUB |LLLLLLLLLL.LLLLLLLLLLLLLLL..LLLLLLLLLLLLLLLLLLL.....LLLLL|
2...,....43...,....44...,....45...,....46...,....47...,....4
AA |SVILDYDTDVSLEKIQPITQNGQHPT|
PHD | |
Rel |62103467766588999999898999
detail:
prH-|000000001112110000000000
prE-|134432111100000000000000
prL-|754456777777888999999898999
subset: SUB |L.....LLLLLLLLLLLLLLLLLL|

END

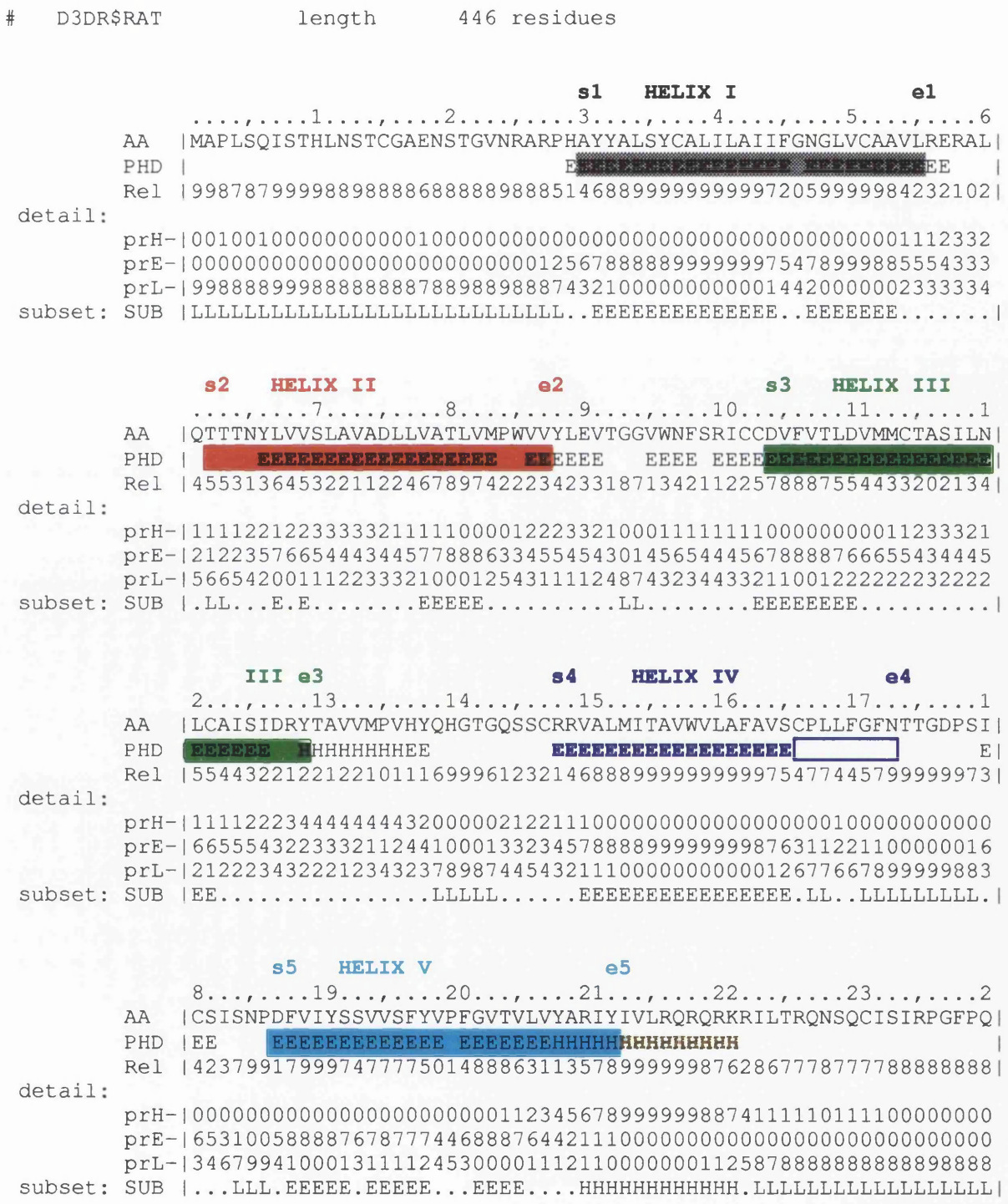
Figure 5.3.6 Profile Network Prediction Heidelberg - PHD; a secondary prediction tool - here applied to the primary structure of the D_2 receptor. Abbreviations: as listed in figure 5.3.5 (page 91). Putative transmembrane helices corespond to those suggested by Baldwin (1993). It is clear that PHD predicts mainly β-sheet for each transmembrane segment. However, PHD predicts α-helix secondary strucutre at the each end of the third intracellular loop (iIII). Also, PHD predicts the short carboxy tail to be mostly α-helix.

```
# d2_human              length      443 residues

                                                  s1      HELIX I             el
                  ....,....1....,....2....,....3....,....4....,....5....,....6
          AA     |MDPLNLSWYDDDLERQNWSRPFNGSDGKADRPHYNYYATLLTLLIAVIVFGNVLVCMAVS|
          PHD    |                                 EEEEEEEEEEEEEEEE  EEEEEEEEE|
          Rel    |997887888999898888888888899889853125677889999999513589998632|
 detail:
          prH-   |001101000000000000000000000000000000000000000000000000001122
          prE-   |000000000000000000000000000000002345677788899999874378999875 4
          prL-   |998888888999898898888888889988887644211100000000025521000002 3
 subset:  SUB    |LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL...EEEEEEEEEEEEEE..EEEEEEE..|

                        s2     HELIX II          e2               s3   HELIX III
                  ....,....7....,....8....,....9....,....10...,....11...,....1
          AA     |REKALQTTTNYLIVSLAVADLLVATLVMPWVVYLEVVGEWKFSRIHCDIFVTLDVMMCTA|
          PHD    |EE        EEEE EE  EEEEEEEEE   HHHHH  EEE   EEEEEEEEEEEEEEEE|
          Rel    |311205763236231110012454775322111211131201211357888754223331|
 detail:
          prH-   |123442112221334334332111101113344443211111111111100000111123
          prE-   |532210012347554443334566776532233333334544334567888766555553
          prL-   |233346775421011122222212012354422122454344543210000122333323
 subset:  SUB    |.....LLL...E.........E.EEE..................EEEEEEE.......|

                    HELIX III   e3                s4    HELIX IV          e4
                  2...,....13...,....14...,....15...,....16...,....17...,....1
          AA     |SILNLCAISIDRYTAVAMPMLYNTRYSSKRRVTVMISIVWVLSFTISCPLLFGLNNADQN|
          PHD    |EEEEEEEEEE  EEEEHHHH          EEEEEEEEEEEEEEEEE             |
          Rel    |312455442121111011121367545522135778999999987323564558998998|
 detail:
          prH-   |232211112222333433443210111133211000000000000000011100000000
          prE-   |544566564333444443212221111111245678889999999886532122200000 01
          prL-   |222221222343212133333478666654322110000000001335675668998998
 subset:  SUB    |....EE................LLL.LL....EEEEEEEEEEEEE...LL.LLLLLLLLL|

                         s5     HELIX V                e5
                  8...,....19...,....20...,....21...,....22...,....23...,....2
          AA     |ECIIANPAFVVYSSIVSFYVPFIVTLLVYIKIYIVLRRRRKRVNTKRSSRAFRAHLRAPL|
          PHD    |   E    EEEEEEEEEEE  EEEEEHHHHHHHHHHHHHHHHHH                |
          Rel    |411227917899756766411366622456799999998776357889878767668888|
 detail:
          prH-   |000000000000000000000112356678899999988776210001100111100000
          prE-   |344541047899777777644677753221000000000000000000000000000000
          prL-   |644458851000121111254210001111100000001112368889888878778888
 subset:  SUB    |.....LL.EEEEEEEEEE....EEE...HHHHHHHHHHHHHH.LLLLLLLLLLLLLLLLL|
```

```
             4...,....25...,....26...,....27...,....28...,....29...,....3
        AA   |KGNCTHPEDMKLCTVIMKSNGSFPVNRRRVEAARRAQELEMEMLSSTSPPERTRYSPIPP|
        PHD  |              EEE                                           |
        Rel  |88678899986311111467977787778987755665444436899988898689877|
detail:
        prH- |00110000000000000011010001110001111111222221000000000100001
        prE- |00000000001344444210001000000000000001001111000000000000000
        prL- |88788999887644444677887888888888777776666578899988898889888
subset: SUB  |LLLLLLLLLLL.......LLLLLLLLLLLLLLLLLLL.....LLLLLLLLLLLLLLLLL|


             0...,....31...,....32...,....33...,....34...,....35...,....3
        AA   |SHHQLTLPDPSHHGLHSTPDSPAKPEKNGHAKDHPKIAKIFEIQTMPNGKTRTSLKTMSR|
        PHD  |                                                        HHH|
        Rel  |78776788899888678898898879788789998855224334479888766431111 2|
detail:
        prH- |10111100000000111000000010100100000012331232200000001234444 4
        prE- |00011000000000000000000000000000000001111001000011110000011 1
        prL- |88877889998888788898898888888889999876556656688888777655544 3
subset: SUB  |LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.......LLLLLLL.......|


                      s6     HELIX VI            e6          s7  HELIX VII
             6...,....37...,....38...,....39...,....40...,....41...,....4
        AA   |RKLSQQKEKKATQMLAIVLGVFIICWLPFFITHILNIHCDCNIPPVLYSAFTWLGYVNSA|
        PHD  |HHHHHHHHHHHHHHHHEEEEEEEEEEEEEEEEEEE          HHHHHEEEEEEE  |
        Rel  |21111345666542113677889865322232221412998773665312566743347 4|
detail:
        prH- |54445566777654443101000000012322221110001113766543211100000 1
        prE- |10110000000113345777888876554554444233000000111335677765321 2
        prL- |34444322111221111111000022333112223545888885111111101123567 6
subset: SUB  |.......HHHHH.....EEEEEEEEE............LLLLL.HHH...EEEE....L.|


             HELIX VII e7
             2...,....43...,....44...,....45...,....46...,....47...,....4
        AA   |VNPIIYTTFNIEFRKAFLKILHC|
        PHD  |    EEEEE  HHHHHHHHHHH  |
        Rel  |5754654051778999999964 8|
detail:
        prH- |0000020114788999999720
        prE- |2116766410000000000000
        prL- |6872112465111000000268
subset: SUB  |LLL.EE..L.HHHHHHHHHHH.L|
```

END

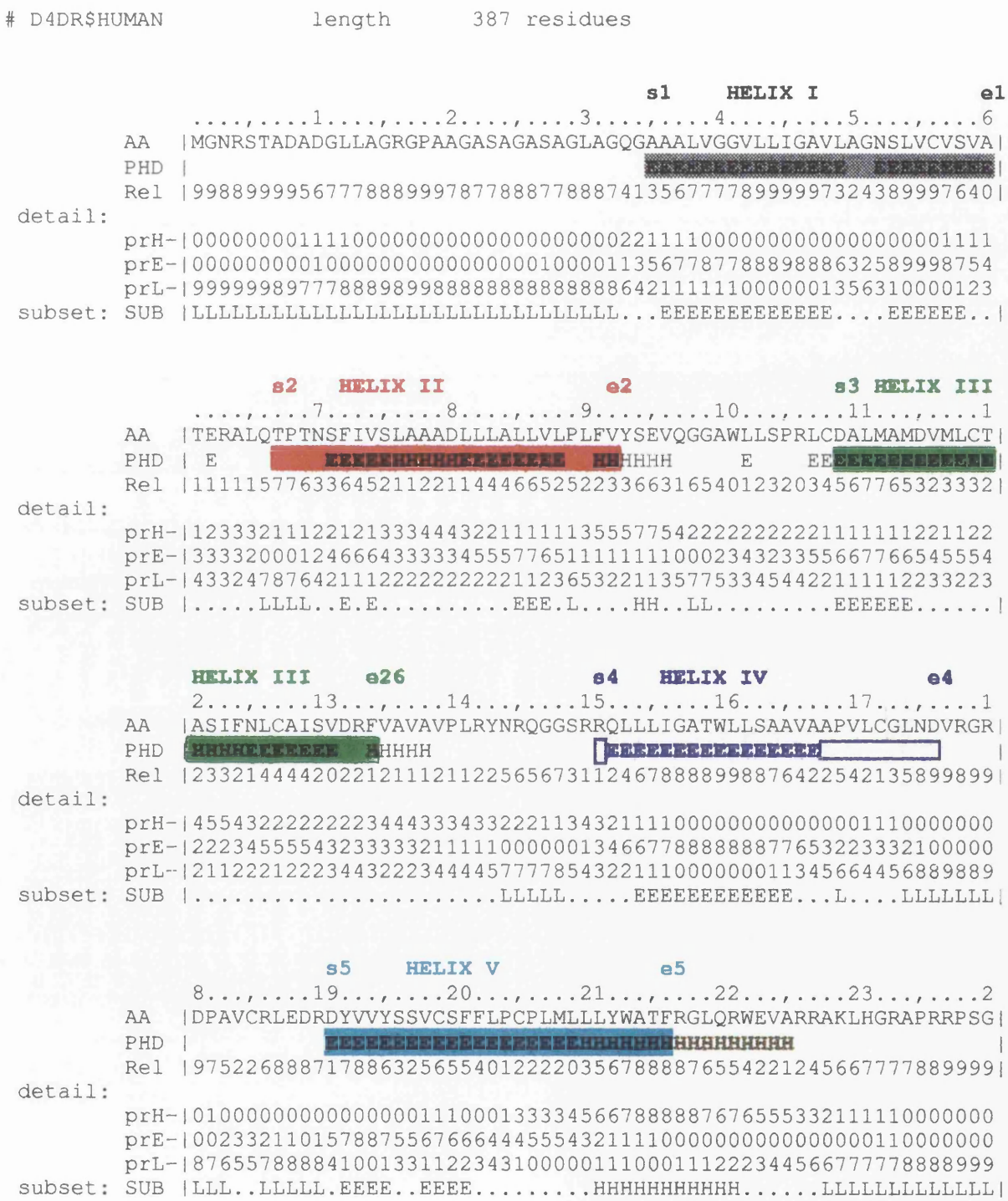
Figure 5.3.7 Profile Network Prediction Heidelberg - PHD; a secondary prediction tool - here applied to the primary structure of the D_3 receptor. Legend same as for D_2 (i.e. figure 5.3.6; pages 93-94).



```
             4..,....25..,....26..,....27..,....28..,....29..,....3
        AA   |QSSCLRLHPIRQFSIRARFLSDATGQMEHIEDKQYPQKCQDPLLSHLQPPSPGQTHGGLK|
        PHD  |                                                          |
        Rel  |8887766654533443442478899889878888888786788777889988888878866|
detail:
        prH- |0000000011132123223210000000010000000001100111000000000000011
        prE- |0000011111111111111100000000000000000000000000000000000000011
        prL- |8888887766655655665688899888888988888887888888889998888888877
subset: SUB  |LLLLLLLLL.L.........LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL|


             0...,....31..,....32..,....33..,....34..,....35..,....3
        AA   |RYYSICQDTALRHPSLEGGAGMSPVERTRNSLSPTMAPKLSLEVRKLSNGRLSTSLRLGP|
        PHD  |                                                          |
        Rel  |7677656655656677788778766879888798889875555446689986431134 87|
detail:
        prH- |1111100122110111000110001110001100000001111221100000124332 11
        prE- |0000121000011100000000010000000000000001111100000001211100 00
        prL- |7788777777777788888888877889888899889887766667889987654556 87
subset: SUB  |LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL..LLLLLLL......LL|


                          s6      HELIX VI          e6          s7 HELIX VII
             6...,....37..,....38..,....39..,....40..,....41..,....4
        AA   |LQPRGVPLREKKATQMVVIVLGAFIVCWLPFFLTHVLNTHCQACHVSPELYRATTWLGYV|
        PHD  |    HHHHHHHHHHHHHHEEEEEEEEEEEEEE HHHHHHH           HHHEEEEEEEEE |
        Rel  |64110233455653202358879998752100134443146889774421334677752 4|
detail:
        prH- |1244455556666544332100000000112345555542101011254322211110 00
        prE- |0011101121011223456888899887543332222211000010013455677776 53
        prL- |7643433222212231111001000012333222211245788877622211000112 36
subset: SUB  |L........HHHH.....EEEEEEEEEE............LLLLLL.......EEEEE..|


             HELIX VII     e7
             2...,....43..,....44..,....45..,....46..,....47..,....4
        AA   |NSALNPVIYTTFNVEFRKAFLKILSC|
        PHD  |      EEEE    HHHHHHHHHHHH  |
        Rel  |5745863543163789999999638 |
detail:
        prH- |0010001110113888998999730 |
        prE- |2122115665310000000000000 |
        prL- |7766873223475110000000268 |
subset: SUB  |LL.LLL.E...L.HHHHHHHHHHH.L|
```

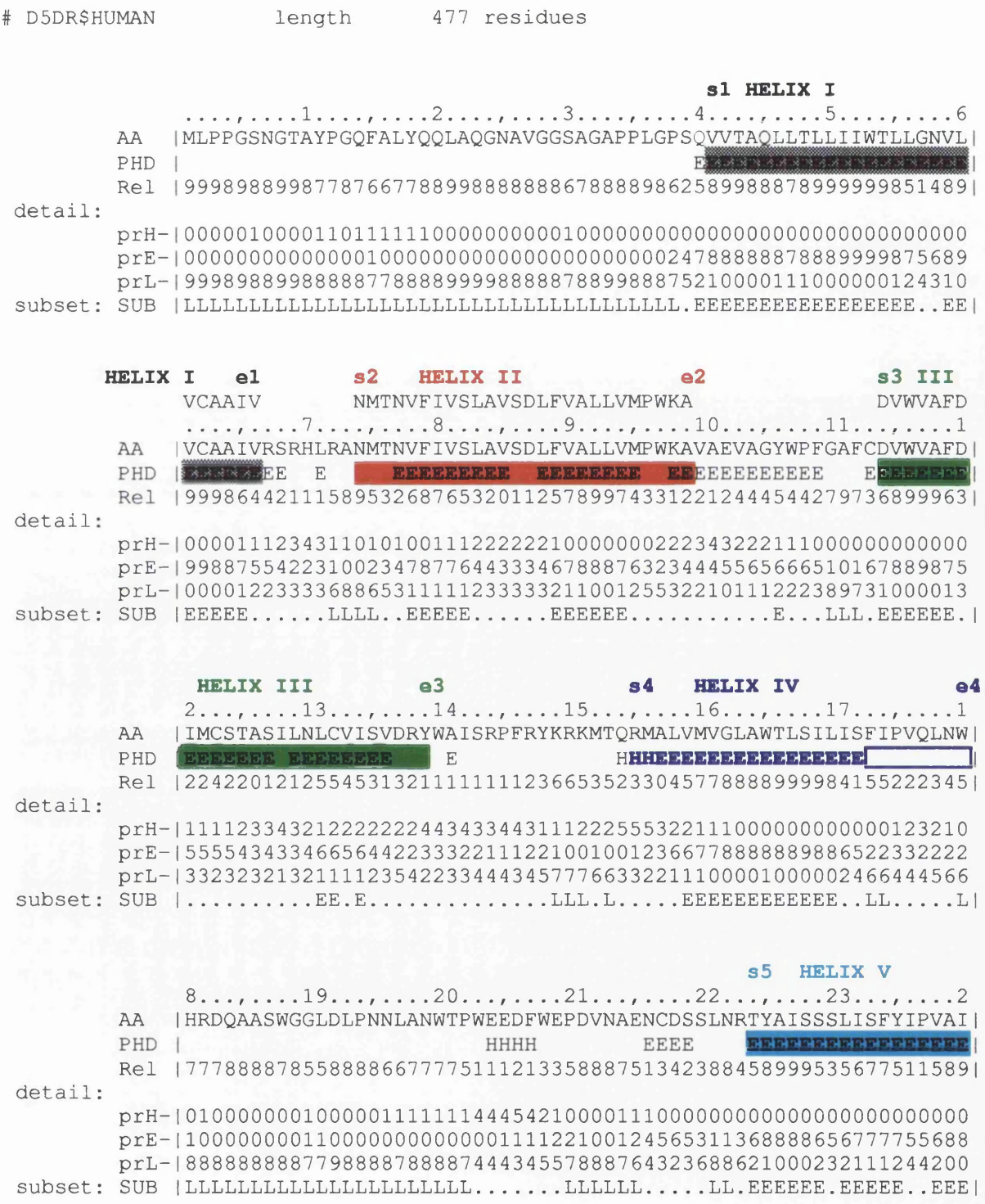
END

Figure 5.3.8 Profile Network Prediction Heidelberg - PHD; a secondary prediction tool - here applied to the primary structure of the D_4 receptor. Legend same as for D_2 (i.e. figure 5.3.6; pages 93-94).



```
                4...,....25...,....26...,....27...,....28...,....29...,....3
          AA    |PGPPSPTPPAPRLPQDPCGPDCAPPAPGLPPDPCGSNCAPPDAVRAAALPPQTPPQTRRR|
          PHD   |                                                            |
          Rel   |998998888887788889988888888888888888888745677888889999987651
detail:
          prH-  |000000000001100000000000000000000011000012211100000000000111
          prE-  |000000000000000000000000000000000000000000000000000000000011
          prL-  |888888888887788888888888888888888888888866788888889999888876
subset:   SUB   |LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.LLLLLLLLLLLLLLLLL|


                                s6   HELIX VI              e6          s1  HELIX VII
                0...,....31...,....32...,....33...,....34...,....35...,....3
          AA    |RRAKITGRERKAMRVLPVVVGAFLLCWTPFFVVHITQALCPACSVPPRLVSAVTWLGYVN|
          PHD   |       HHHHHHHHHHEEEEEEEEEEEE EHHHHHHH          EEEEEEEEE   |
          Rel   |44576502346654322466678775322113444422699987610234466653135|
detail:
          prH-  |22211245567765553211110000011234555555210000113332222111011
          prE-  |11000001111112335677778877554343222211210000001245666766542
          prL-  |66787643321121111111111123333222211234789888743211001123456
subset:   SUB   |..LLLL....HHH.....EEEEEEEE.............LLLLLLL......EEEE...L|


                HELIX VII   e7
                6...,....37...,....38...,....39...,....40...,....41...,....4
          AA    |SALNPVIYTVFNAEFRNVFRKALRACC|
          PHD   |      EEE    HHHHHHHHHHHHH  |
          Rel   |667971331452678999999997379
detail:
          prH-  |11001111123778999999997510
          prE-  |11101455421000000000001000
          prL-  |77787432356511100000001389
subset:   SUB   |LLLLL.....L.HHHHHHHHHHHH.LL|



END
```

Figure 5.3.9 Profile Network Prediction Heidelberg - PHD; a secondary prediction tool - here applied to the primary structure of the D_5 receptor. Legend otherwise the same as for D_1 (i.e. figure 5.3.5; pages 91-92).



```
                 V       e5                                                                  s6 VI
                 4...,....25...,....26...,....27...,....28...,....29...,....3
          AA     |MIVTYTRIYRIAQVQIRRISSLERAAEHAQSCRSSAACAPDTSLRASIKKETKVLKTLSV|
          PHD    |EEEEHHHHHHHHHHHHHHH                                      HHHHHHHHHHHHE|
          Rel    |975222568999998665176679898887877888888899754744568998764 22|
detail:
          prH-   |012345678899998776511110000000000000000000001112667888877643
          prE-   |886532111000000000000000000000000000000000002100011000011235
          prL-   |001112110000001122487789999888888888899989986686211100011111
subset:   SUB    |EEE...HHHHHHHHHHHH.LLLLLLLLLLLLLLLLLLLLLLL.L..HHHHHHHH...|

                   HELIX VI          e6                        s7      HELIX VII
                 0...,....31...,....32...,....33...,....34...,....35...,....3
          AA     |IMGVFVCCWLPFFILNCMVPFCSGHPEGPPAGFPCVSETTFDVFVWFGWANSSLNPVIYA|
          PHD    |EEEEEEEEEE  HHHHHHHHHE                 HHHHHHEEEEEE       EE  |
          Rel    |5457898752002233202211489999998874113775322566653147468422 8|
detail:
          prH-   |211000000023445554454310000000001234577654321110110110011100
          prE-   |666788877532333234333320000000000011110122566776542121024531
          prL-   |122100012333211111212258999999988654211122011112346767863358
subset:   SUB    |E.EEEEEEE.............LLLLLLLLLL....HHH...EEEEE...L.LL....L|

                 VII e7
                 6...,....37...,....38...,....39...,....40...,....41...,....4
          AA     |FNADFQKVFAQLLGCSHFCSRTPVETVNISNELISYNQDIVFHKEIAAAYIHMMPNAVTP|
          PHD    |  HHHHHHHHHHHHHHH                                    EEE      |
          Rel    |92788999999998615578767543435666557777888888777532442799999 |
detail:
          prH-   |048889999999987521101111111111111001000000010111000000000000
          prE-   |000000000000000000000101222210000110000000000001256631000000
          prL-   |851100000000011477888777666677777788888888888865323579999 9
subset:   SUB    |L.HHHHHHHHHHHHH.LLLLLLLL....LLLLLLLLLLLLLLLLLLL.....LLLLLLL|

                 2...,....43...,....44...,....45...,....46...,....47...,....4
          AA     |GNREVDNDEEEGPFDRMFQIYQTSPDGDPVAESVWELDCEGEISLDKITPFTPNGFH|
          PHD    |                                                          |
          Rel    |9878898898899787555456899999877777335678762124478878788 9|
detail:
          prH-   |001000000000000121122100000000111110001011100010001000000
          prE-   |000000000000000001110000000001000123210001343221000010000
          prL-   |987889889889988866666789999987788756778887545667888888899
subset:   SUB    |LLLLLLLLLLLLLLLLLLL.LLLLLLLLLLLLLL..LLLLLL.....LLLLLLLLLL|


END
```

5.5 Discussion

5.5.1 Homology Analysis

It is clear from the Needleman and Wunsch (1970) alignments (table 5.3.1) and the random shuffling alignments (tables 5.3.2 to 5.3.5) that there is no homology between bRh and GPCRs. Therefore to adopt homology modelling based strategies in which models of a particular GPCR are based exactly on bRh is questionable. Indeed, Henderson (personal communication: appendix 2) has suggested that bRh is not a good starting model for modelling GPCRs. A low resolution 2D projection structure of rhodopsin (rH) is available (Gebhard *et al.*, 1993). Given that rhodopsin is a coupled to a G protein and has significant homology with the DA receptors (table 5.3.2) it follows that the structural data relating to rH should take precedence over bRh. Hence Baldwin's (1993) paper on the probable arrangement of transmembrane helices in GPCRs which is based on the work of Gebhard *et al.* (1993) should form the basis of all GPCR modelling studies.

The calculation of Z scores provides the basis to classify dopamine receptors in terms of whether they are a member of the D_1 or D_2 sub-type dopamine GPCRs. Equally importantly, the calculation of Z scores provides the basis for rapidly classifying any future cloned GPCR. Since whole sequences were used to calculate Z scores then by default both TM homology (associated with the ligand binding site) and homology relating to the G protein-coupling regions is used in the computation. This is important as Donnelly *et al.* (1994) has recently shown that the evolution of GPCRs occurs at two sites: the ligand binding site and the G protein-coupling regions.

5.5.1.1 PEPPLOT Analysis of Bacteriorhodopsin (bRh)

The sponsor wanted bRh to be used as a template to model the dopamine D_2 GPCR and so it deserves some attention. Acidic and basic residues appear to be clustered at the beginning and end of hydrophobic stretches of residues. It is clear that the KD plot finds five very clear peaks and two less clear hydrophobic peaks. The GES plot fails to

detect the third and fourth transmembrane helices. (The structure of bRh has been solved and the hydophobicity peaks are known to corespond to the actual location of transmembrane helices in the integral membrane protein - e.g. Jähnig, 1990). It is clear that β-sheet is more favoured than α-helix - this simply supports the argument that secondary prediction algorithms are not good at predicting secondary structure in integral membrane proteins. It is clear that turns in the transmembrane region tend to occur at the helix start and end postions in aggreement with the findings of Green, 1990.

5.5.1.2 PEPPLOT Analysis of D_1 and D_2

D_1 (like D_5) has a short third intracellular loop between transmembrane helices V and VI. This is reflected in the hydropathy analysis where there is a gap between the prominent peaks coresponding to helices V and VI. Transmembrane helix does not show up clearly in the hydropathy analysis indicating that it is more hydrophilic than the other transmembrane helices. Also, the D_1 subfamily of DA receptors is characterised by a long intracellular carboxy tail. D_2 (like D_3 and D_4) has a long third intracellular loop between transmembrane helices V and VI and a shorter tail.

5.5.2 How Do The KD and GES plots of the Dopamine Receptors compare to bRh?

It is clear that there is a definite similarity in the plots between bRh and both D_1 and D_2 sub-families (figure 5.3.4). This suggests that the overall topology of the dopamine family of GPCRs is similar to that of bRh. That is, the transmembrane region of the dopamin receptors consists of seven transmembrane helices.

5.5.3 Tripeptide Glycoslylation Sequences in the Dopamine Family of GPCRs

Glycosylated sites are invariably found exposed to the exterior medium - Cotmore *et al.*, 1977. Also, tripeptide sequences found on exposed loops or β-turns are likely to be glycosylated. The distribution of potential tripeptide glycosylation sequences in the

dopamine primary structures (table 5.3.6) suggests that the amino-tail is found on the extracellular side of the membrane like that of bRh. The loop connecting transmembrane helices IV and V in D_3 is likely to be extracellular given that it has a potential tripeptide glycosylation sequence. Given that the carboxy-trail lacks potential glycosylation sites - this suggests it is found on the intracellular side of the membrane - again like the short carboxy-tail of bRh.

5.5.4 Profile Network Prediction Heidelberg - PHD

The neural network (PHD - Rost and Sander, 1993; Rost *et al.*, 1994) has been trained using globular proteins. However, the rules governing folding of the non-membrane component of large integral membrane proteins like DA GPCRs are likely to be the same as for globular proteins. PHD was applied to each member of the dopamine family of GPCRs to predict α-helix (H), β-strand (E) and loop (L); pages 91 to 100. It is clear that PHD has a tendency to predict β-sheet in those regions of the primary structures that corespond to the transmembrane region[1] of the integral membrane dopamine receptors. However, for every dopamine receptor sub-type, PHD predicts to 80% accuracy α-helix secondary structure at the beginning third intracellular loop (iIII) and a α-helix secondary structure at the end the third intracellular loop (between 50% and 100% accuracy) and is in aggreement with the findings of MaloneyHuss and Lybrand (1992) who also predicted a short α-helix at both end of iIII. In addition, the short carboxy tail, which terminates with a conserved Cys residue, in D_2, D_3 and D_4 (i.e. the entire D_2 sub-family) are predicted at the 90 to 100% accuracy to be α-helix. Also, PHD predicts at the 90 to 100% accuracy level that there are three turns of α-helix between the C-terminal end of ptms VII and the first Cys residue in the long carboxy tail of the D_1 sub-family (D_1 and D_5). It has been suggested that the Cys residue in catechol amine binding GPCRs (and others) is palmitated and anchored to the membrane. For example, the Swissprot entry for β_2-adrenergic receptor (entry: P07550) states that Cys-346 (the first Cys residue upstream of TM7) is palmitated. Also, Ross (1989) showed that the amino-terminus segment of the carboxyl terminal (carboxy tail) have a co-operative role in anchoring GPCRs to the

[1] Based on the Joyce Baldwin (1993) assignments of transmembrane helices in GPCRs.

membrane. It follows that the short α-helix probably runs along the surface of the membrane. This would require a 90^0 turn immediately following TM7. There is a clear precedence for such a turn in integral membrane proteins. The X-ray crystal structure of integral membrane protein Light Harvesting Complex II from *Pseudomonas acidophilia* (strain 10050) is in the process of being solved here in Glasgow. At the N-terminal end of the repeating α-helix subunit there is a sharp 90^0 turn allowing the polypeptide chain to negotiate the surface of the lipid membrane (McDermott *et al.*, submitted, 1994).

5.6 Conclusion

Homology analysis suggests that the use of bRh as a starting template to model the dopamine family of GPCRs is flawed. There is not sufficient homolgy present to justify homology modelling of any GPCR using bRh as a template. In contrast it is clear that rH exhibits significant homolgy with each member of the dopamine family of GPCRs. Hence, any structural data related to rH such as the projection map produced by Gebhard *et al.* (1993) is of direct use in modelling the dopamine family of GPCRs. However, the hydropathy analysis did suggest that the transmembrane region of the dopamine family of GPCRs exhibits the same overall topology as bRh. That is, a hepta-helical motif characterises the transmembrane region of GPCRs. Secondary structure prediction algorithms are not reliable when applied to the transmembrane regions of integral membrane proteins such as bRh and GPCRs. However, secondary structure prediction algorithms such as PHD can be usefully applied to the non-transmembrane regions of GPCRs.

6. Detection of Regions of Internal Homology in the Dopamine Family of G Protein Coupled Receptors

6.1 Summary

The N and/or C-terminal ends of the third intracellular loop of G protein coupled receptors frequently contain G protein-activator sequences (Nishimoto and Okamoto, 1992; Ikezu *et al.*, 1992). Repeat sequences have been reported in the middle of the third intracellular loop in the D_4 subtype dopamine receptor (Van-Tol *et al.*, 1992). Here we report the frequent occurrence of regions with palindromic-like character and short stretches of straight and reverse repeat sequences in the dopamine family of receptors (D_1 through to D_5). Such *homology events* are found particularly in the amino terminal region, putative transmembrane regions and in the third intracellular loop of each dopamine subtype. Their frequency of occurrence suggests multiple roles in the establishment and functioning of the dopamine receptor. To our knowledge, this is the first report of palindromic-like sequences in the catecholamine receptor family.

6.2 INTRODUCTION

The dopamine family of catechol amine binding G protein coupled receptors (GPCRs) play a central role in the development of such disorders as rigidity in Parkinson's disease and hallucinations in schizophrenia and Alzheimer's disease (reviewed in chapters 1 to 4). While multiple sequence alignments suggest an important role for several residues (e.g. hydrogen bonding of serines in the amino-half of putative transmembrane helix V to the dopamine ligand). Such alignments do not reveal regions of internal homology. The motivation to perform an internal homology study of the dopamine family of receptors was to see if it could be used to detect helix end points. The hypothesis being that regions of internal homology are unlikely to cross conformational boundaries such as from transmembrane helix to random coil.

6.2.1 Types of Internal Homology

There are four types of internal homology studied here: straight repeats, reverse repeats, palindromic regions and *tube sequences* (table 6.2.1). Reverse repeats are the same as straight repeats except they involve reading the sequence forwards and also backwards whereas straight repeats are detected by reading the same sequence in the forward direction only. A palindromic sequences differs from a reverse repeat in that the reverse homology applies only to the same area of the primary structure and exhibit obvious symmetry and where they occur within helices they result in a pseudo-twofold axis in their structure (Suzuki, 1992). *Tube sequences* are similar to palindromic sequences except that the *tube* sequence does not exhibit complete symmetry - and for this reason can be considered *snake sequences* as they have an obvious head or tail - they do not read exactly the same way in both directions.

Sequence	Type
248 265 PPAPRLPQDPCGPDCAPP PPAPGLPPDPCGSNCAPP _____X__X_____XX_____ 264 281	Straight repeat - top numbers indicate start and end position of sequence in the complete primary structure (re: first line of amino-acid residues); likewise for bottom numbers (re: 2nd line of amino-acid residues). Underscores indicate hits and Xs indicate miss. (From D_4)
202 211 FLPCPLMLLL FLPLVLLALL ____XX_XX___ 91 82	Reverse repeat. From D_4
157 163 SIVWVLS SLVWVIS _____	Short Palindromic sequence. The second line of residues is the top sequence reversed. The sequence thus reads the same in both directions - represented by an absence of capital Xs. From D_2
37 50 ALVGGVLLIGAVLA ALVAGILLVGGVLA ___X_____X___	Tube sequence. The second line of residues is the top sequence reversed. While the tube sequence is palindromic-like it does look (read) differently from each end. Hence its similarity to a London tube train or snake. (let V = I) - from D_4

Table 6.2.1. Explanation and description of types of internal homology observed in this study

6.3 EXPERIMENTAL

Two simple algorithms to detect: Internal Homology (IH) and Reverse Homology (RH) were coded in 3L Parallel FORTRAN (version 2.1.2) - appendix 3 for code listing. Three primary structures with known internal homology (sero transferrin - McGillivary *et al.*, 1983; lactotransferrin - Powell and Ogden, 1990; ovotransferrin - Williams *et al.*, 1982) were used as *positive* controls to help debug and test the IH Scan algorithm. Using a sliding window internal homology was detected by scanning each sequence and calculating pairwise alignment. Typically, for a window length of 10, any consecutive series of scores of 6 or more (corresponding to a stringency of 60%) were investigated further by careful examination of the primary structure to decide the exact nature of the internal homology.

6.4 RESULTS

The IH and RH Scan Algorithms was applied to each subtype of the dopamine family of receptors and muscarinic M_3 and M_5 (tables 6.4.1 to 6.4.6; pages 108-116). Sample IH output (re: D_1 and D_4) is displayed graphically in figure 6.4.1 (page 117). Figure 6.4.2 (page 118) illustrates the IH results in graphical form with regard to the positive controls with known internal homology.

Sequence (D_1)	Type	Location	Comments
43 50 LVCAAVIR LFCATLIR _X__XX__ 30 23	Reverse Repeat	N-terminal end	
72 76 LVAVL LVAVL _____	Palindromic	Helix II	
88 97 GFWPFGSFCN GFWVFVDFTN ___X_XX_X_ 320 311	Reverse Repeat	iII Helix VII	
152 159 163 VLISFIPVQLSW VPIFSILVSLTW _X_XX_X_X___ 159 152 148	Reverse Repeat/Tube	Helix IV	Reverse repeat includes a tube sequence: 152-159 (S = T)
211 217 IVTYTRI IRTYTVI _X___X_	Tube	Helix V	
229 235 AALERAA LFCATLI __X_X__	Tube	iIII	
284 294 CWLPFFILNC CNLIFFPLWC _X_X__X_X_	Tube	Helix VI	Very similar to tube in helix VI in D_5
313 319 FDVFVWF FWVFVDF _X___X_	Tube	Helix VII	
420 432 LSVILDYDTDVSL LSVDTDYDLIVSL ___XX___XX___	Tube	C-terminal end	

Table 6.4.1 Regions of internal homology in human D_1 subtype dopamine receptor (Dearry *et al.*, 1990). Palindromic-like or tube sequences together with straight and reverse repeat sequences are shown. Helix assignments are those given in SWISS-PROT accessed using the GCG package (Devereux *et al.*, 1984).

Sequence (D_2)	Type	Location	Comments
38 46 ATLLTLLIA AILLTLLTA _X_____X_	Tube	Helix I (1st half)	
52 59 NVLVCMAV NYLMPMAV _X_XX___ 143 136	Reverse Repeat	Helix I (2nd half) iII	
91 97 VVYLEVV VVELYVV __X_X__	Tube	Helix II (2nd half)	
107 113 CDIFVTL CIIFVGL _X___X_ 385 379	Reverse Repeat	Helix III (1st Half) Helix VI (1st half)	
157 163 SIVWVLS SLVWVIS _____	Palindromic	Helix IV	(L = I)
189 198 FVVYSSIVSF FSVISSYVVF _X_X__X_X_	Tube	Helix V (1st half)	See next entry
206 216 LLVYIKIYIVL LVIYIKIYVLL _X_____X_	Tube	Helix V (2nd half)	Includes the palindromic sequence: YIKIY or VYIKIYI (V = I). Proline residue occupies middle position in helix V.
271 278 EAARRAQE EQARRAAE _X____X_	Tube	iIII	Located in middle of iIII

Table 6.4.2 Regions of internal homology in human D_2 subtype dopamine receptor (Dal-Toso *et al.*, 1989). Palindromic-like or tube sequences together with straight and reverse repeat sequences are shown. Helix assignments are those given in SWISS-PROT accessed using the GCG package (Devereux *et al.*, 1984). [Continued overleaf].

Sequence (D_2 - continued)	Type	Location	Comments
311 317 SHHGLHS SHLGHHS __X_X__	Tube	iIII	Located in middle of iIII
338 343 AKIFEI AKRFEI __X___ 436 431	Reverse Repeat	iIII iIV	
354 364 SLKTMSRRKLS SLKRRSMTKLS ___XX_XX___	Tube	iIII	
416 425 YVNSAVNPII YVVFAPNAII __XX_X_X__ 192 183	Reverse Repeat	Helix VII Helix V	

Sequence (D_3)	Type	Location	Comments
38 42 ALILA ALILA _____	Short palindromic sequence	Helix I	
72 81 AVADLLVATL AFALVWVATI _X_XXX____ 163 154	Reverse repeat	Helix II Helix IV	(I = L)
191 198 YSSVVSFY YFSVVSSY _X____X_	Tube	Helix V	
213 223 IVLRQRQRKRI IRKRQRQRLVI _XX_____XX_	Tube	iIII	iIII: 210 - 374; hence this sequence marks the start of the N-terminal side of iIII
283 299 LLSHLQPPSPGQTHGGL LGGHTQGPSPPQLHSLL _XX_X_X___X_X_XX_	Tube	iIII	Middle of iIII.
309 316 TALRHPSL TSLRGNSL _X__XX__ 93 86	Reverse Repeat	iIII eII	
331 341 SLSPTMAPKLS SLKPAMTPSLS __X_X_X_X__	Tube	iIII	2nd half of iIII.
347 361 LSNGRLSTSLRLGPL LPGLRLSTSLRGNSL _XXX_____XXX_	Tube	iIII	sequence marks end of C-terminal side of iIII; contains the palindromic sequence: RLSTSLR
390 397 PFFLTHVL PMVLTAVL _XX__X__ 84 77	Reverse Repeat	Helix VI Helix II	
410 417 LYRATTWL LYNTTTQL __XX__X_ 67 60	Reverse Repeat	Helix VII Helix II	

Table 6.4.3 Regions of internal homology in rat D_3 subtype dopamine receptor (Sokoloff *et al.*, 1990). Palindromic-like or tube sequences together with straight and reverse repeat sequences are shown. Helix assignments are those given in SWISS-PROT accessed using the GCG package (Devereux *et al.*, 1984).

Sequence (D_4)	Type	Location	Comments
19 31 AAGASAGASAGLA ALGASAGASAGAA _X_____X_	Tube	N-terminal end	Includes the palindromic sequence GASAGASAG
37 50 ALVGGVLLIGAVLA ALVAGILLVGGVLA ___X_____X___	Tube	Helix I	Pseudo-two fold rotational symmetry (I = V)
43 47 LLIGA LLIGA _____ 154 158	Straight Repeat	Helix I Helix IV	
60 66 ATERALQ ATFRGLQ __X_X__ 214 220	Straight Repeat	iI iIII	
96 103 VQGGAWLL VVVGAFLL _XX__X__ 318 325	Straight Repeat	eII Helix VI	
98 104 GGAWLLS GATWLLS _XX____ 157 163	Straight Repeat	eII Helix IV	
161 171 LLSAAVAAPVL LVPAAVAASLL _XX_____XX_	Tube	Helix IV	Pseudo-two fold rotational symmetry
202 211 FLPCPLMLLL FLPLVLLALL ___XX_XX__ 91 82	Reverse Repeat	Helix V Helix II	
209 215 LLLYWAT LLLIGAT ___XX__ 153 159	Straight Repeat	Helix V Helix IV	

Table 6.4.4 Regions of internal homology in human D_4 subtype dopamine receptor (Van Tol *et al.*, 1991). Palindromic-like or tube sequences together with straight and reverse repeat sequences are shown. Helix assignments are those given in SWISS-PROT accessed using the GCG package (Devereux *et al.*, 1984). [Continued overleaf].

Sequence (D_4 - continued)	Type	Location	Comments
226 233 RRAKLHGR RRAKITGR _____X__ 301 308	Straight Repeat	iIII iIII	Putative low-potency non-selective G protein-activator and selective G_i/G_o-activator respectively. (L = I).
234 250 APRRPSGPGPPSPTPPA APPTPSPPGPGSPRRPA __XX__X___X__XX__	Tube	iIII	This tube marks the start of the famous repeat sequence of D4.2 subtype (Van-Tol, *et al.*, 1992).
248 265 PPAPRLPQDPCGPDCAPP PPAPGLPPDPCGSNCAPP ____X__X____XX____ 264 281	Straight Repeat	iIII	Middle of iIII; referred to as D4.2 subtype (Van-Tol, *et al.*, 1992. See figure 6.4.1
264 273 PPAPGLPPDP PPGPGSPRRP __X__X_XX_ 244 235	Reverse Repeat	iIII iIII	
298 302 RRRRR _____	Poly R Sequence	end of iIII	Polyarguinine sequence.
316 325 LPVVVGAFLL LPLVLLALLL __X_XX_X__ 90 81	Reverse Repeat	Helix VI Helix II	Helix II also has reverse homology with Helix V.

Sequence (D_5)	Type	Location	Comments
16 22 ALYQQLA ALQQYLA __X_X__	Tube	N-terminal end	
30 39 SAGAPPLGPS SPGLPPAGAS _X_X__X_X_	Tube	N-terminal end	Helix I: 40 - 66; hence this tube sequence marks the start of helix I (Swissprot file)
46 56 LLTLLIIWTLL LLTWIILLTLL ___X___X___	Tube	Helix I	Middle of helix I. (I = L)
83 92 LAVSDLFVAL LAVFLDSVAL ___XXXX___	Tube	Helix II	
88 95 LFVALLVM LTWALGVM _XX__X__ 167 160	Reverse Repeat	Helix II Helix IV	
168 172 SILIS SILIS _____	Palindromic	Helix IV	
227 233 ISSSLIS IFSILIS _X_X___ 174 168	Reverse Repeat	Helix V Helix IV	
250 258 RIAQVQIRR RRIQVQAIR _XX___XX_	Tube	iIII	iIII: 247 - 296; since this tube occurs at the beginning of iIII and contains the motif BB (where BB are two basic residues) this tube forms part of a putative low-potency non-selective G protein-activator sequence (Nishimoto *et al.*, 1992). Using neural network this tube sequence is predicted to be a helix to greater than 82% accuracy.

Table 6.4.5 Regions of internal homology in human D_5 subtype dopamine receptor (Sunahara *et al.*, 1991). Palindromic-like or tube sequences together with straight and reverse repeat sequences are shown. Helix assignments are those given in SWISS-PROT accessed using the GCG package (Devereux *et al.*, 1984). [Continued overleaf].

Sequence (D_5 - continued)	Type	Location	Comments
269 277 AQSCRSSAA AASSRCSQA _X_X_X_X_	Tube	iIII	An alternating tube sequence
306 319 VCCWLPFFILNCMV VMCNLIFFPLWCCV _X_X_X__X_X_X_	Tube	Helix VI	Pseudo alternating tube
326 334 PEGPPAGFP PFGAPPGEP _X_X_X_X_	Tube	eIII	An alternating tube sequence
341 347 FDVFVWF FWVFVDF _X___X_	Tube	Helix VII	
424 430 EVDNDEE EEDNDVE _X___X_	Tube	C-terminal end	
457 465 LDLEGEISL LSIEGELDL _X_____X_	Tube	C-terminal end	(I = L)

Sequence (Muscarinic)	Type	Location	Comments
30 46 VITIAVVTAVVSLITIV VITILSVVATVVAITIV ____XX_X_X_XX____	Tube	Helix I	M_5; helix I: 30 - 53.
154 163 AWLISFILWA AWLIFSILWA ____XX____	Tube	Helix IV	M_5; helix IV: 147 - 169. This tube sequence is also found in M_1.
162 170 ISFDRYFSI ISFYRDFSI ___X_X___	Tube	End of Helix III	M_3; DRY motif plays an important role in G protein activation.

Table 6.4.6 Some tube sequences found in human M_3 and M_5 subtype muscarinic acetylcholine receptors (Peralta *et al.*, 1987; Bonner, *et al.*, 1988). Helix assignments are those given in SWISS-PROT accessed using the GCG package (Devereux *et al.*, 1984).

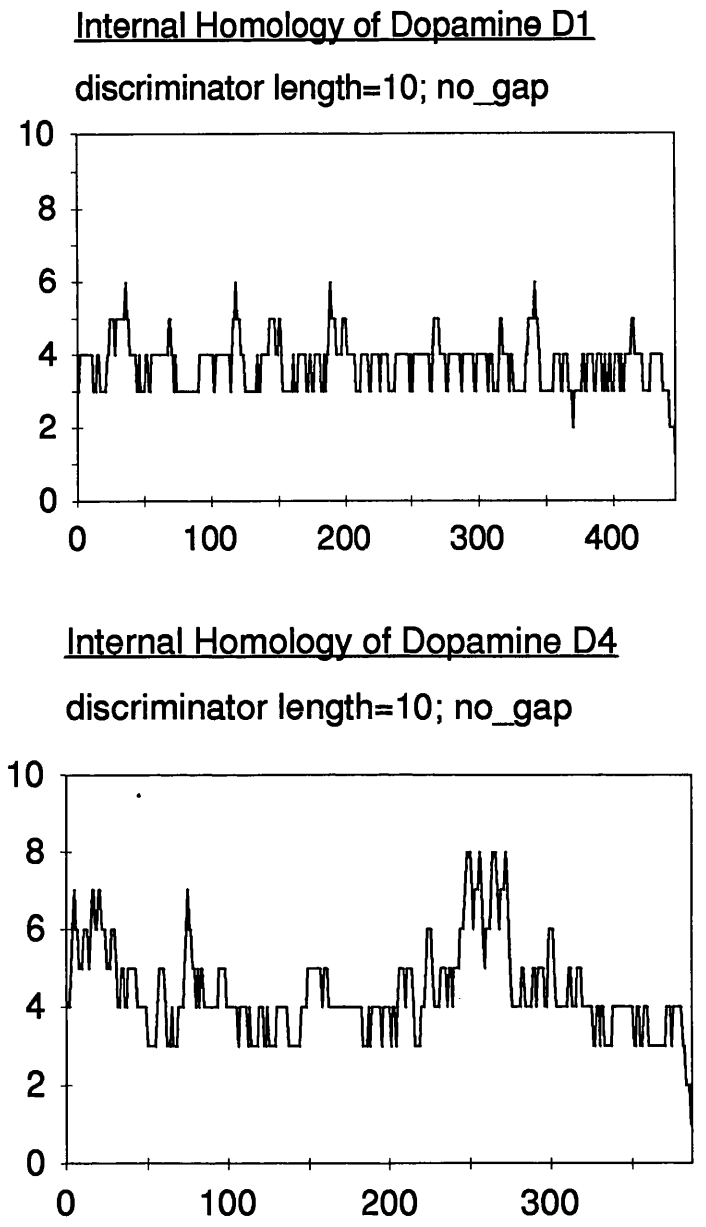


Figure 6.4.1 IH (Internal Homology) Scan algorithm, which detects repeat sequences, applied to human D_1 (top) and human D_4 (bottom) subtype DA receptors. It is clear that D_4 has more repeat homolgy than D_1 - the reason for this is not clear. Also, the famous repeat sequence in the third intracellular loop (iIII) is depicted as two identical peaks: residues 248-265 is repeated at 264-281; see table 6.4.1. This repeat sequence was observed by the author prior to the publication describing the possible significance of repeat regions in iIII (Van Tol *et al.* 1992).

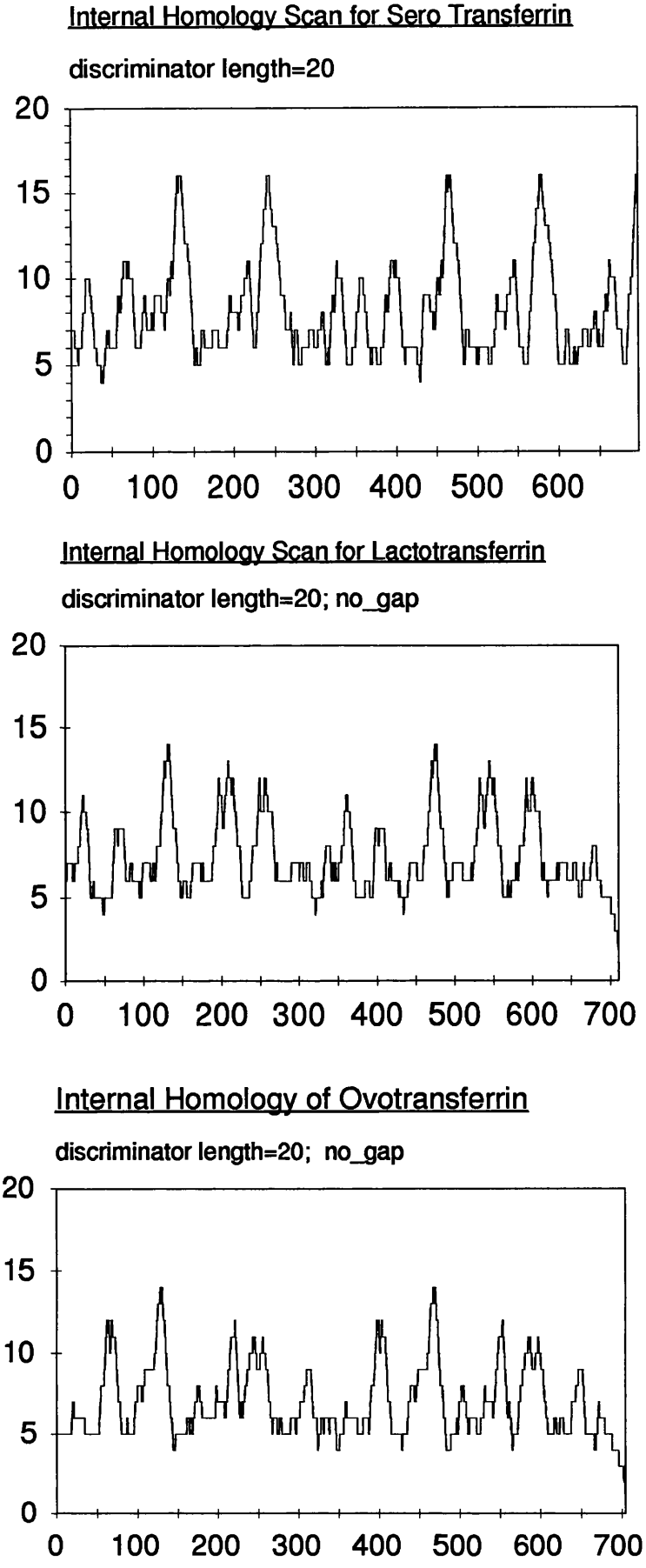


Figure 6.4.2 Internal Homology (IH) Scans applied to the primary structures of sero transferrin (McGillivary *et al.*, 1983), lactotransferrin (Powell and Ogden, 1990) and ovotransferrin (Williams *et al.*, 1982). It is clear from the symmetry of each plot that there are two distinct structural lobes. These *positive* controls were used to debug the algorithm.

6.5 DISCUSSION

IH and RH detect straight repeats and reverse repeats respectively. Subtype D_4 receptor appears to be quite remarkable in terms of the frequency and distribution of tube sequences - the reasons for this are not clear. We have also run RH on two other cationic amine GPCRs: M_3 and M_5 subtype muscarinic acetylcholine receptors (table 6.4.6) and found evidence of tube sequences. Unfortunately it was not possible to predict the location of helix end points any better than predictions based on simple hydropathy analysis. However, it is clear that tube sequences are well represented in putative helical secondary structures. Possible reasons for this are discussed below.

6.5.1 Direction markers in protein folding

Unlike true palindromic sequences, tube sequences include a sense of direction - somewhat like a snake - the unfortunate person who accidentally steps on a poisonous snake looks immediately at the head of a snake to assess the best direction of escape! This follows from the simple observation that tube sequences do not read the same in both directions. For example, the tube sequence in helix I of the D_4 dopamine subtype receptor (residues 37 to 50: ALVGGVLLIGAVLA) has a glycine at position 4 and an alanine at position 11 (I = V); hence this tube does not read exactly the same in both directions. A tube train on the London Underground to the casual observer looks identical viewed from both ends accept that a driver is located at one end only. G protein coupled receptors, like all multihelix transmembrane proteins, are believed to fold in a two stage process (Bormann *et al.*, 1992). Each α-helix is folded independently across the bilayer and then the helices are assembled into a tertiary structure in which the helices are not much altered. Tube sequences may aid the assembly resulting in the tertiary structure. In a very recent paper, Lemmon *et al.* (1994) claim that specific helix-helix interactions inside lipid bilayers guide the folding process. They report a pattern of 7 amino acids (LIxxGVxxGVxxT) which when introduced into several hydrophobic transmembrane α-helices promotes their specific dimerization. They point out that since this motif is rare, whilst specific helix association is not, many other such motifs may exist, which could guide folding and oligomerization. It is

quite possible that the tube sequences identified here are the "many other such motifs" hypothesised by Lemmon and co-workers. The pattern was discerned in previously reported mutagenesis and computational modelling studies of specific dimerization of the single transmembrane α-helix of human glycophorin A (Lemmon *et al.*, 1992a, 1992b). Fourier analysis (chapter 8) designed to detect periodicity in α-helices (and β-sheets) could be used to detect the kind of periodicity apparent in the pattern of 7 amino acids described by Lemmon and co-workers. That is, the residues responsible for the specific interactions occur at every third or fourth position characteristic of an α-helix with conserved residues on one face of the helix. Fourier analysis (chapter 8) clearly show that several of the transmembrane helices of the hepta-helical motif of DA GPCRs also exhibit periodicity consistent with residues being conserved on one face of the helix.

6.5.2 Harmonic oscillators

It has been suggested that the occurrence of polymorphic variation in the form of variable numbers of repeat sequences in the middle of iIII (third intracellular loop) of the dopamine D_4 receptor subtype may play an important role in susceptibility to neuropsychiatric disease and in responsiveness to antipsychotic medication (Van-Tol, *et al.*, 1992). Unless polymorphic variation is demonstrated for a particular tube sequence in a receptor subtype casting such a role can not be predicted for tube sequences. However, tube sequence may play an important role influencing the molecular dynamics in secondary structures. For example, tube sequences occur in two separate halves of helix V of the D_2 dopamine subtype receptor (table 6.4.2); a conserved proline occupies the middle position separating these tube sequences. It is quite possible that each tube sequence in each half of helix V act as harmonic oscillators. Molecular dynamics simulations of helices with a proline occupying the middle position with particular reference to transmembrane helices in non-polar environments have been carried out (Sankararamakrishnan *et al.*, 1991). It was shown that such helices oscillate between straight structures and bent structures on a pico-second time scale. The location of tube sequences in opposite halves of helices containing a proline occupying the middle position was not considered.

6.5.3 Molecular recognition

The possibility that tube sequences may play an important role in molecular recognition is based on a number of recent findings. For example, it has already been noted that basic-domain sequences which are almost palindromic or self-symmetrical play a role in molecular recognition (Suzuki, 1992). The sequence marking the start of iIII and end of iIII of the dopamine family of receptors are also dominated by basic residues and act as G protein-activator regions (Nishimoto and Okamoto, 1992; Ikezu *et al.*, 1992). We have established that the start and end sequences of iIII in each member of the dopamine family of receptors probably adopt helical structures (using a trained neural network (Bukhard and Sander, 1993). Using the neural network the tube sequence at the start of iIII of the D_5 subtype dopamine receptor (250-258; see table 6.4.5) is predicted to from part of a helix to greater than 82% probability. Likewise the tube sequence at the start of iIII of D_3 subtype dopamine receptor (213-223; see table 6.4.3) is predicted to form part of a helix to greater than 82% accuracy and is dominated by basic residues. These finding are in line with others (Maloney-Huss and Lybrand, 1992) who have also predicted that the start and end sequences of iIII of β_2-adrenergic receptor adopt an α-helical structure. Hence it is possible that this tube sequence incorporated in a helical structure plays a vital part in stimulating G proteins.

6.5.4 Origin of Tube Sequences

The possibility that tube sequences might in fact have originated from palindromic sequences was considered by examining the DNA sequence coding tube sequences of at least 10 residues in length. If tube sequences originated from palindromic sequences then this should be clearly expressed in terms of single base pair changes in those codons specifically responsible for differentiating a tube sequence from a palindromic sequence. Only in two cases (figures 6.5.4.1 and 6.5.4.2) is it clear that a single base-pair change could be responsible for converting what might have been a palindromic sequence into a tube sequence.

```
37                                                          50
  A   L   V   G   G   V   L   L   I   G   A   V   L   A
  A   L   V   A   G   I   L   L   V   G   G   V   L   A
          X                                   X
  _   _   _   _   _   _   _   _   _   _   _   _   _   _


GCG CTG GTG GGG GGC GTG CTG CTC ATC GGC GCG GTG CTC GCG
GCG CTC GTG GCG GGC ATC CTC CTG GTG GGC GGG GTG CTG GCG
___ ___ ___  Z  ___ ___ ___ ___ ___ ___  Z  ___ ___ ___
```

Figure 6.5.4.1 Tube sequence in D_4 starting at residue 37 and ending at position 50. The single codon responsible for differentiating this tube sequence from a palindromic sequence is a single base difference (represented by a capital Z). If the cytosine was converted to guanine then this tube would be a palindromic sequence. It is possible in this instance that a single base mutation occurred (C to G or G to C) in what was originally a palindromic sequence to generate a tube sequence. (V = I).

```
154                                                     163
  A   W   L   I   S   F   I   L   W   A
  A   W   L   I   F   S   I   L   W   A
                  X   X
  _   _   _   _   _   _   _   _   _   _


GCC TGG CTG ATC TCC TTC ATC CTC TGG GCC
GCC TGG CTC ATC TTC TCC ATC CTG TGG GCC
___ ___ ___ ___  Z   Z  ___ ___ ___ ___
```

Figure 6.5.4.2 Tube sequence in M_5 starting at residue 154 and ending at position 163. The single codon responsible for differentiating this tube sequence from a palindromic sequence is a single base difference (represented by a capital Z). It is possible in this instance that a single base mutation occurred (C to T or T to C) in what was originally a palindromic sequence to generate a tube sequence.

6.6 CONCLUSION

It has been said that at some point we are likely to encounter a truly unexpected surprise among membrane proteins (Richardson and Richardson, 1989). The distribution of homology events throughout the dopamine family of GPCRs is one such surprise. For the sake of rational drug design and the treatment of many serious medical disorders linked to the dysfunction of specific GPCRs it is vital that the structural and functional reasons for such surprises are solved. At the very least, unravelling the meaning of such surprises will enrich our understanding of the biology of the supergene class of GPCRs. It is just unfortunate that IH and RH scans are not able to detect helix end points.

7. DH Scan - a novel alignment method to detect putative transmembrane segments of G protein coupled receptors using massive parallel processing on a transputer machine.

7.1 Summary

A novel sequence alignment method has been developed to exploit the computing power offered by massive parallel processing (MPP). DH Scan uses the characteristic topology of G-protein coupled receptors and discriminator homology to reveal the receptors overall topology (including the seventh putative transmembrane helix which is not easily discernible in hydropathy plots) and to provide an independent verification of the likely starting points of transmembrane spanners. DH Scans of members of the dopamine (DA) family of receptors show that DH Scan can be used as a classification tool and has a role in suggesting chimeric genetic analysis experiments. DH Scan is designed to run automatically on anything from 2 transputers, up to several hundred transputers. The algorithm offers increased performance with every additional transputer.

7.2 Introduction

G protein coupled receptors (GPCRs) are a major class of transmembrane proteins found in multicellular eukaryotic organisms (Findlay *et al.*, 1990). They play an essential role in the transfer of information across the membrane (Findlay, 1991). The topology of GPCRs is thought to consist of seven transmembrane helices similar to those of bacteriorhodopsin (bRh; Henderson and Unwin, 1975; Henderson *et al.*, 1990). While sequence alignments indicate that bRh is not related to GPCRs, hydropathy analysis indicates that their polypeptide chains exhibit the same folding pattern (e.g. Fasman and Gilbert, 1990). High transmembrane homology has been noted in the G-protein coupled receptor families - for example, the dopamine (DA) family (Sibley and Monsama, 1992). Hence identification of regions of high transmembrane homology may provide a basis for locating helix end positions.

7.2.1 Alignment Methods

Currently, the main alignment methods used to compare two GPCRs are based on the Needleman and Wunsch (1970) algorithm. Multiple sequences can be compared using methods suggested by for example Corpet (1988) using data processors (von Heijne, 1991). However, approaches that calculate the different degrees of sequence conservation are rare. As Friemann and Schmitz (1992) pointed out, one method used to mark conserved sequence sections is to box identical residues found in aligned columns or to present conserved amino acids as asterisks. Such methods only give a visual impression of the positions of clustered sequence identities. In the Friemann and Schmitz approach to this problem, protein sequences must be aligned before their algorithm can be used to display identities and differences. In contrast, DH Scan clearly displays sections of sequence conservation when starting off with unaligned protein sequences.

Attwood *et al.* (1991) have developed a novel alignment method based on pattern-matching discriminators compiled specifically to suit the characteristic topology offered by the GPCRs. This method has been used to locate transmembrane regions along the protein sequences of distantly related GPCRs. To work satisfactorily, this method requires precise details concerning the nature and position of each amino acid along each transmembrane protein sequence. Weighting factors must also be computed and inputted into the program to ensure suitable matches between transmembrane regions of distantly related GPCRs. Also, the algorithm does not cope explicitly with the possibility of gaps (insertions and deletions). In contrast, DH Scan is explicitly coded to cope with gaps and deletions and does not require any computation of weighting factors.

7.2.2 Application of Parallel Processing

Argos (1987) developed a sensitive procedure to compare amino acid sequences by comparing every possible span of length L residues in one protein with all such spans in the second protein. The alignment algorithm was later parallized and mounted on a transputer based machine (Vogt and Argos, 1992) and is used to align distantly related sequences in large databases. In our approach, we have explicitly set out to align G-coupled protein

receptors and to exploit the characteristic topology offered by this class of transmembrane proteins. We anticipate that our algorithm will be further developed to detect regions of high functionality. For example, stretches of amino acids involved in binding with G-proteins on the cytoplasmic side of the protein and the identification of the amino acids responsible for molecular recognition on the extracellular surface of the protein.

The Third International Conference on Applications of Transputers (1991) cited 141 papers. Only two of these papers were devoted to the application of transputers in the biosciences (Gonzalez and Lopez, 1991; Kulkarni *et al.*, 1991). This suggests that less than 2% of papers on applications of transputers have any direct connection with the biosciences. Also, Vogt and Argos (1992) have reported that prior to their publication on the use of transputers in searching distantly related protein sequences, the use of transputers in molecular simulations and secondary structure prediction constituted the very few published examples relating to molecular biology known to them (cf. Raine *et al.*, 1989; Boehncke *et al.*, 1990). The only additional published work known to us also describes the use of transputers in molecular biology simulations (cf. Goodfellow, 1990; Goodfellow *et al.*, 1990; Jones and Goodfellow, 1990).

The exact reasons for the slow uptake of parallel technology based on transputers in the biosciences is not clear. However, the case for using this technology is clear, for example Vogt and Argos (1992) reported that a parallel implementation of their sensitive sequence alignment algorithm using a Macintosh IIc host computer and 21 transputers achieved 22 times the speed of a VAX 8650 at a fraction of the cost. Consequently, this chapter places considerable emphasis on describing the operation and role of parallel computing technology.

7.2.3 Nuts and Bolts of Parallel Processing

In the traditional Von Neumann Architecture digital computer operations were performed sequentially - an instruction was fetched and decoded, the operands (the values to be operated on - if any) fetched, the operation executed, and the result stored. None of

these operations were started until the preceding one was complete. The designers of parallel computers based on the transputer have diverged from the strict Von Neumann Architecture by allowing transputers to work on different parts of the same algorithm simultaneously, passing messages from transputer to transputer according to the needs of the algorithm. However, as Vogt and Argos (1992) point out, not all algorithms are amenable to parallelisation. We were fortunate in that DH Scan is extremely amenable to parallelisation. In fact, we found that DH Scan is best run as a "processor farm" application (Parallel FORTRAN User Guide, 1991; Tregidgo and Downton, 1990).

7.2.4 Processor Farms

Building a parallel application normally requires detailed knowledge of the transputer network on which the application is intended to run. For many parallel computations it is useful to be able to create applications which will automatically configure themselves to run on any network of transputers. Processor farm applications will automatically run faster when more transputers are added to a network, without recompilation or reconfiguration. Junk (1991) has pointed out that farming is the simplest form of transputer-based parallelism. Rea (1991) has also noted that "task-farming" is an extremely efficient form of parallelism being as much as 30% more efficient than algorithmic parallelism.

In the processor farm technique, an application is coded as one master task which breaks the job down into small, independent pieces called work packets which are processed by separate worker tasks. All the worker tasks run the same code which is automatically copied across an arbitrary network of transputers. Work packets are automatically distributed to idle worker tasks by routing software supplied with the compiler. Each worker simply accepts work packets, processes them, and sends back result packets via the same routing software. A worker task is simply a list of sequential code designed to carry out certain basic functions: wait for initialisation data (if any); read the initialisation data; wait for a work packet, read the work packet; process the work packet; send back a result packet; repeat until the main program has finished.

7.2.5 Parallel Architectures

Though in principle parallel computers based on arrays of transputers are MIMD machines, that is: multiple instruction multiple data or "message passing" machines (Dan Charles, 1992), when in processor farm mode, transputer based parallel computers resemble SIMD, that is: single instruction multiple data - an approach sometimes called "data parallelism". In the SIMD approach (pioneered by ICL through their Distributed Array Processor (DAP machines; Reddaway, 1992), data is distributed among the processors, which then march in step through a set of instructions, carrying out the same operations on different pieces of data. As Vogt and Argos (1992) point out, the most efficient implementation of sequence alignment algorithms use SIMD. In fact, Jones (1992) has recently reported a sequence pattern matching method which has been implemented on a massively parallel computer - the 8192 processor CM-2 manufactured by Thinking Machines Inc. SIMD machines achieve higher computational speeds by consecutively doubling the number of processors. The CM-2 machine used by Jones (1992) has 8192 processors (2 to the power 13) - the next machine along is the 16384 processor CM-2, that is: 2 to the power 14 processors. Scaling with processor farms is guaranteed by simply adding one or more transputers to the network at marginal financial cost.

7.3 Systems and methods

7.3.1 Hardware - Minimum Requirements

The current version of DH Scan assumes the following general configuration: a host computer, an IBM compatible PC (MSDOS version 3.0 or later) with a hard disk and at least 512 kbytes of RAM, coupled to a parallel machine made up of T800 or T805 transputers. The T800 is rated at 1.9 Mflop (12.5 mips at 25 MHz clock speed); the T805 is rated at 4.3 Mflop (30 mips at 30 MHz clock speed). While the T800 and T805 each have four serial communication channels (figure 7.3.1.1) or links to permit flexible architectures - DH Scan will run on a simple linear chain of transputers (figure 7.3.1.2). An

IMS B004 (pronounced *boff*) compatible transputer board with 8MB of RAM must be fitted to the host computer. This plug-in card handles the communication between the PC and the parallel computer and its single T800 transputer provides additional processing power. A standard mono-monitor is required to allow the program to interact with the user. DH Scan has been run using an array of 32 T800 transputers partitioned into 8 and 24 transputers. Hence with the host B004 this makes for : 9 and 25 transputers respectively. But DH Scan will run on any network of transputers - see next section.

7.3.2 Hardware - How Many Transputers?

DH Scan will compile and run without modification on a chain of transputers ranging from 2 transputers up to several hundred transputers. For example, a DH Scan of human D_2 (443 amino acids in length) with rat dopamine D_3 will run most efficiently on an array made up of 443 T805 transputers - one transputer for each amino acid position of D_2; this run would take around 2 minutes whereas the same run on a single T800 transputer (equivalent to an Intel 486DX processor) would take around 70 hours. A large array of Inmos T9000 transputers (an order of magnitude faster than the T800) would enable DH Scans to be carried out in an "interactive" fashion. Scaling is achieved by simply adding/using more transputers. For example: 1, 2, 4, 8, 16, 32, 64, 128, 256 T800 transputers would take approximately: 70, 35, 17, 8, 4, 2, 1, 0.5, 0.25 hours respectively to run DH Scan using D_2 and D_3 as input files.

DH Scan can be implemented on "supernodes" of 1000 or more transputers - but some transputers would not receive a work packet and so would remain idle. With minor modification, DH Scan could fully utilise a supernode - an extra couple of lines of code to split work packets up into smaller work packets would suffice. However, every work packet requires two messages - a work packet sent from the master task to an idle worker task and a returning results packet. Since the communication overhead can become significant, the maximum practical number of transputers would be of the order of several hundreds rather than thousands. The overall objective being to keep each transputer busy for a period considerably greater than the time taken to receive and send each data packet. This problem is known as the grain size problem (Tabak, 1990).

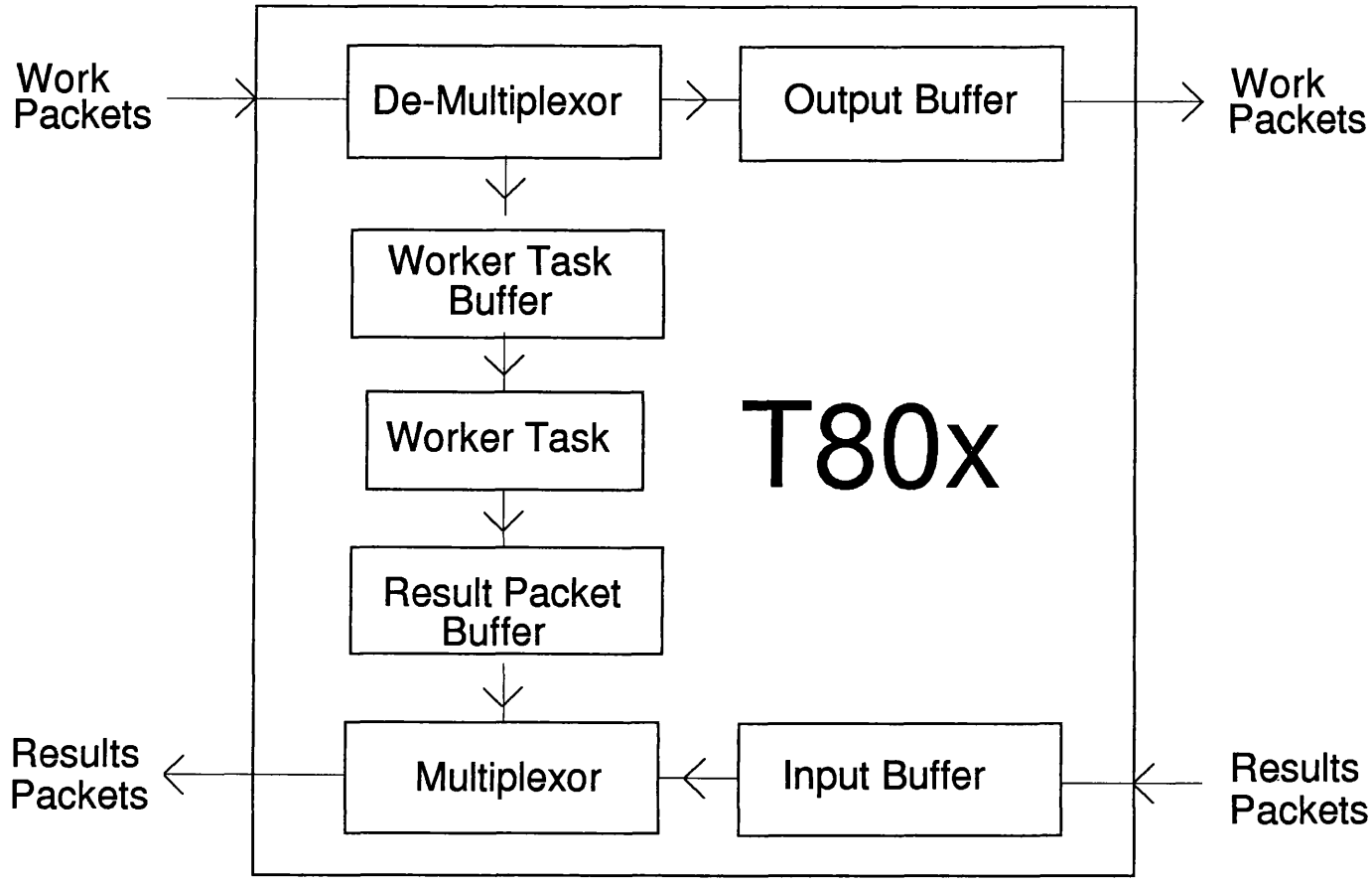


Figure 7.3.1.1 Schematic representation of data packet flow through a T800/T805 transputer located in a linear tree. The T800 and T805 each have four serial communication channels or links to permit flexible architectures. The de-multiplexing process selects a route for incoming work packets, either to the worker task process via the worker task buffer, or to the next transputer in the chain via the output buffer. Work packets are only buffered along the outgoing leg of the tree, the F77_NET_SEND command of the main thread of the master task is blocked when all the buffers are saturated and all the worker tasks are busy. Studies by Rea (1991) indicates that the work packet buffer need only store a single work packet to speed up the processing by up to 25% prior to saturation. A work packet buffer just one work packet "thick" would guarantee equitable load balancing throughout the tree. The result packet buffer allows the worker task to become receptive to accepting a new work packet without having to wait on the multiplexor. The multiplexor interleaves the outgoing result packet with the stream of incoming results packets being received from transputers further down the tree, which are collectively forwarded on to the receive thread of the master task.

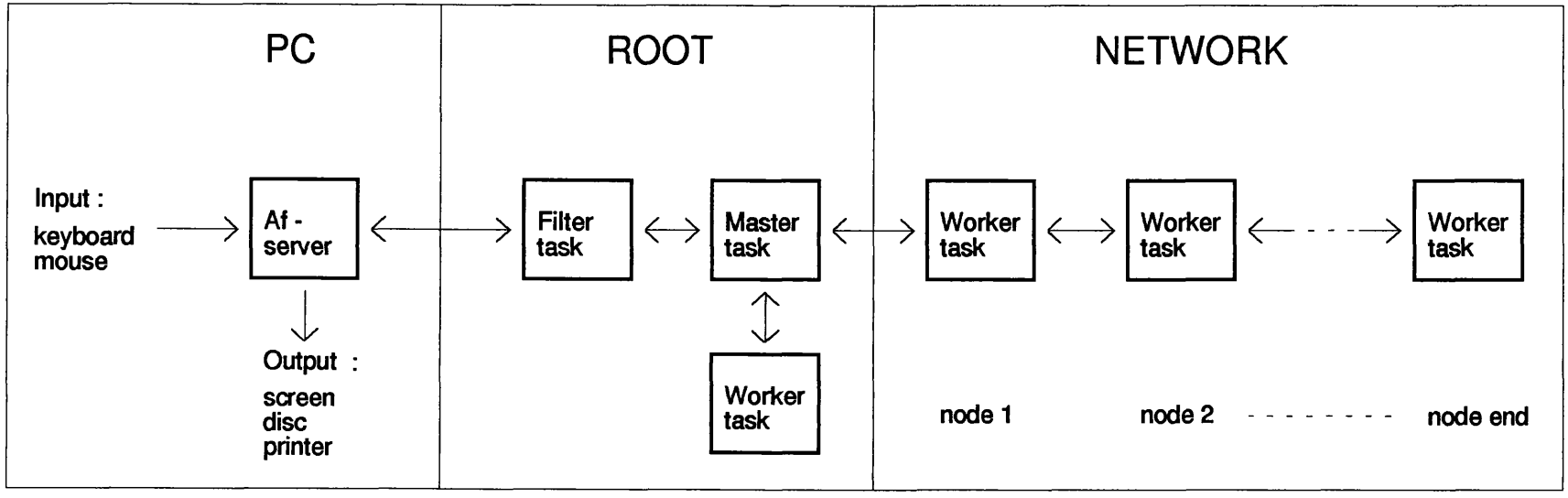


Figure 7.3.1.2 Processor farm application is made up of : the MSDOS alien file server (af server) program which executes on the PC or host computer ; filter task and master task which resides on the B004 board, which collectively are known as the root software ; worker tasks - one worker task is copied onto every transputer (node) of the network. In some implementations of the processor farm approach an extra worker task resides on the root transputer as shown.

7.3.3 Hardware - Future Trends

Parsytec of Aachen, the German manufacturer of parallel systems, has already announced its T9000 based GC or "Grand Challenge" architecture - Parallelogram, (1991); Allan (1992). These machines will offer performance in the 1 to 400 GigaFLOPS range (i.e. greater than 10 to the 11 floating point operations per second). GC systems are built up from one or more GigaCube modules. Each GigaCube module comprises a three dimensional array of 64 transputers. This 3-D topology can be expanded, with a scaleable increase in performance, up to 256 GigaCubes. Due to delays in the shipment of the T9000 the first GC machine is being built with T805 transputers. Parsytec are also leading the European TeraFLOP Initiative to build the world's first TeraFLOP machine which will use 65,536 T9000 transputers to create a super-massively parallel MIMD system.

7.3.4 Software

7.3.4.1 Software - Specifics.

DH Scan was written in Parallel FORTRAN (version 2.1.4) which is available from 3L Ltd., Livingston, Scotland. Parallel FORTRAN is an explicit compiler, it does not support automatic parallelization. The internationally accepted standard for FORTRAN (ANSI X3.9-1978 and ISO 1539-1980) is supported by the compiler (Parallel FORTRAN User Guide, 1991). Parallel FORTRAN also supports various useful extensions to the ANSI Standard (commonly referred to as the FORTRAN F77 standard) such as the DO WHILE statement. A number of additional subroutines in the library allow communication of data packets (up to 1024 bytes) around the transputer network. The library also includes something called the DOS Package . The functions and subroutines of the DOS Package allow a program running on a transputer system which is hosted by an MSDOS computer to access the software interrupts, DOS function calls and the memory of the host system.

7.3.4.2 Software - Communication

The user can be prompted to input data in the usual way via the keyboard. When the program is executing on the transputer net, the master task can keep the user informed of progress by writing updates to the screen as results packets arrive from the worker tasks. For obvious reasons it is not advisable to permit the master task to write anything to the screen if the number of worker tasks exceeds about 30. The master task is the only piece of code directly connected to the Inmos alien file server (afserver) and so is permitted access to the FORTRAN run-time input/output (i/o) system (Cooper and Allan, 1992). The afserver program is an MSDOS program which is mounted on the PC and provides communication between the host computer and the B004 board (also known as root). Root software includes the filter task (essential), master task (essential) and a worker task (optional). The filter program is mounted automatically onto root at the time of execution - it passes on messages travelling in both directions between the afserver program and the master task. Worker tasks are automatically copied onto each node (transputer) of the linear network at execution time.

One big advantage offered by the processor farm approach is that the transfer of messages (i.e. data packets) between transputers can be achieved with three simple calls. With regard to 3L Parallel FORTRAN these calls are: F77_Net_Broadcast_Send, F77_Net_Send and F77_Net_Receive. The master task communicates with the worker tasks by using all three calls: the F77_Net_Broadcast_Send function call is used to broadcast initialisation data throughout the network; the F77_Net_Send function call is used to forward work packets to idle worker tasks; the F77_Net_Receive function call is used to receive incoming results packets sent by worker tasks. The worker tasks communicate with the master task by using two of the calls: the F77_Net_Receive function call is used to receive initialisation data and work packets; the F77_Net_Send function call is used to send results packets back to the master task. Restricting all communications to just three function calls makes the coding of processor farm applications very straightforward. In particular, a message passing harness based on Occam-2 is not required.

7.3.4.3 Software - Threads

A particularly useful feature of Parallel FORTRAN is its ability to allow the user to take advantage of the transputer's architecture by creating new concurrent threads of execution within a task. Parallel FORTRAN's threads resemble the "processes" of Modular-2, and the "coroutines" of some other languages. Each thread has its own stack (allocated by its creator), but shares its code with any other threads in the same task. Threads of the same task also have access to the same COMMON blocks. Semaphore functions in the run-time library are used to prevent threads which share data from interfering with each other. However, the multiple thread facility supported by 3L Parallel FORTRAN differs from that supported by 3L Parallel C. Parallel FORTRAN threads are not able to invoke the same subprogram more than once at a time - Parallel FORTRAN subprograms are not reentrant. Since threads of the same task have access to the same COMMON blocks, communication between threads is easily achieved using globally declared variables. In addition, communication can be achieved using specific address channels.

7.3.4.4 Software - Portability

The usual programming language associated with the transputer is Occam 2 (Inmos, 1988). There are serious draw backs in using Occam, not least its lack of portability. 3L Parallel FORTRAN supports the more portable F77 standard. Also, 3L are active in porting its compilers to a wide variety of platforms. For example, 3L have just released Parallel C for the 50 Mflop Texas Instruments C40 DSP (digital signal processing) chip. Parallel machines based on C40 chips are already being manufactured. Given that FORTRAN compilers are very much in demand, it seems reasonable to expect a future releases of 3L Parallel FORTRAN to support the C40 chip. 3L are in fact marketing their compilers as "Multiplatform Parallel Development Tools".

7.3.4.5 Software - Implicit Versus Explicit Compilers

Tabak (1990) identifies two types of parallelization: implicit and explicit parallelization. In explicit parallelization the user must first recognise those parts of the program which can be performed independently of other parts without interfering with the overall flow of the program. The next step is to write the code so that those parts of the program which can be carried out in parallel are allocated to separate processors. In contrast, implicit parallelization can take sequential code and run it on parallel hardware without requiring a major effort to re-write the code. It is the compiler which recognises those parts of the program which do not depend on data generated by other parts of the program - a technique referred to as data flow analysis. Junk (1990) has studied this problem and has concluded that automatic parallelization, capable of distributing an arbitrary program over several transputers to any degree of efficiency, cannot exist for transputers. We have found that replicative applications such as DH Scan run efficiently using explicit parallelization.

Writing explicit code is perceived as being very difficult. However, the processor farm approach greatly simplifies the overall program logic. The program logic is broken down into two sets of code: the master task and the worker task. Concurrent processes in the master task are coded as threads. Usually only three threads (main, send and receive) are required and each of these follow a basic pattern repeated in all processor farm applications. The code for each worker task is identical and so requires a single effort to write and debug it. Similarly, communication between the master task and each worker task usually occurs at certain points in the program logic - a basic pattern repeated in all processor farm applications. A novice parallel programmer has but to read the Mandelbrot Program Listing in the Parallel FORTRAN User Guide (1991) to perceive the overall program logic.

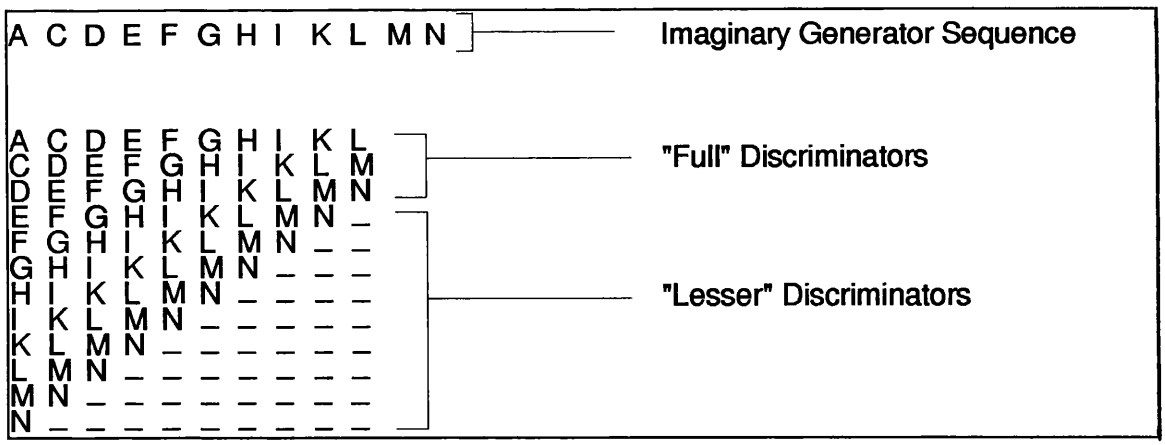


Figure 7.3.5.1; Generation of discriminators of 10 amino acids in length (d=10) from an imaginary generator sequence of 12 amino acids in length (n_{gs}=12). Three (that is, n_{gs}-d+1) "full" discriminators are generated plus nine (that is, d-1) "lesser" discriminators with progressive numbers of non-scoring blanks.

7.3.5 Algorithm

7.3.5.1 Algorithm - General Theory

The program exploits unaligned sequences of homologous GPCRs (for source code listing see appendix 3). Discriminators are generated from one protein sequence (called the generator sequence) and compared to a second protein sequence (called the target sequence), as follows:

$$\text{number of discriminators} = l - d + 1$$

Where l is the length of the generator sequence and d is the discriminator length. However, it is simpler to generate l discriminators. For example, a generator sequence of say length 450 amino acids and a chosen discriminator length of 20 amino acids; 431 (that is, l-d+1) "full" discriminators are generated plus 19 (that is, d-1) "lesser" discriminators with progressive numbers of non-scoring blanks (figure 7.3.5.1). It has been noted by several workers that the non-polar transbilayer helices tend to be around 20 or so amino acids in

length (for example, Engelman *et al.,* 1986); the discriminator length should therefore be in the range 20-25. In our runs we have typically used a discriminator length of 25. For example, a DH Scan of dopamine D1 (446 amino acids in length) against dopamine D5 (477 amino acids in length) and a discriminator length of 25 amino acids would generate 422 full discriminators and 24 lesser discriminators from the generator sequence D1. Each discriminator generated from the generator sequence D1 is moved along the whole length of the target sequence D5 and scores computed on a pair-wise hit/miss basis and the highest score for each discriminator noted. This first part of DH Scan is referred to as the No_Gap_Sweep.

Further sweeps allow for one gap (One_Gap_Sweep), two gaps (Two_Gaps_Sweep), one deletion (One_Del_Sweep) and two deletions (Two_Del_Sweep) - each sweep is carried out in sequence. Highest scores from one sweep are carried forward to the next sweep. Each completed sweep results in two output files in two column format - one file contains the highest discriminator scores with regard to each amino acid position of the generator sequence and the other file contains the highest discriminator scores with regard to each amino acid position of the target sequence. Since there are currently five sweeps (No_Gap_Sweep, One_Gap_Sweep, Two_Gap_Sweep, One_Del_Sweep and Two_Del_Sweep), DH Scan produces ten output files in all. There are plans to incorporate a One_Gap_One_Del_Sweep, where one gap and one deletion are incorporated into every possible position in each discriminator generated from the generator sequence.

The two output files resulting from the One_Gap_Sweep contain the highest scores following the No_Gap_Sweep and One_Gap_Sweep; the two output files resulting from the Two_Gap_Sweep contains the highest scores following the No_Gap_Sweep, One_Gap_Sweep and the Two_Gap_Sweep; the two output files resulting from the One_Del_Sweep contains the highest scores following the No_Gap_Sweep, One_Gap_Sweep, Two_Gap_Sweep and the One_Del_Sweep; the two output files resulting from the Two_Gap_Sweep contains the highest scores following the No_Gap_Sweep, One_Gap_Sweep, Two_Gap_Sweep, One_Del_Sweep and the

Two_Del_Sweep. The AWK utility under UNIX can be used to compare the results of each file.

The One_Gap_Sweep/One_Del_Sweep and the Two_Gap_Sweep/Two_Del_Sweep generate large numbers of discriminators from the generator sequence (considered in detail in section 7.3.5.2.2). As a result, using a single T800 transputer to sequentially perform a complete No_Gap_Sweep of D1 against D5, takes around 3 minutes whereas a complete One_Gap_Sweep takes 1 hour and 20 minutes and a full DH Scan around 70 hours. It is clearly important to remember that each discriminator is moved along the entire length of the target sequence resulting in significant numbers of pair-wise hit/miss comparisons.

7.3.5.2 Algorithm - Processor Farm Approach

DH Scan has been designed to run as a processor farm application (see figure 7.3.5.2). The application actually consists of two tasks: (i) align_m.f77 - this is the master task, and runs in the root transputer; (ii) align_w.f77 - this is the worker task, and runs in all the other transputers of the net (figure 7.3.1.2). It is also possible to run a worker task on the root transputer to mop up any spare capacity. It follows that it is also possible to run DH Scan on just the root transputer.

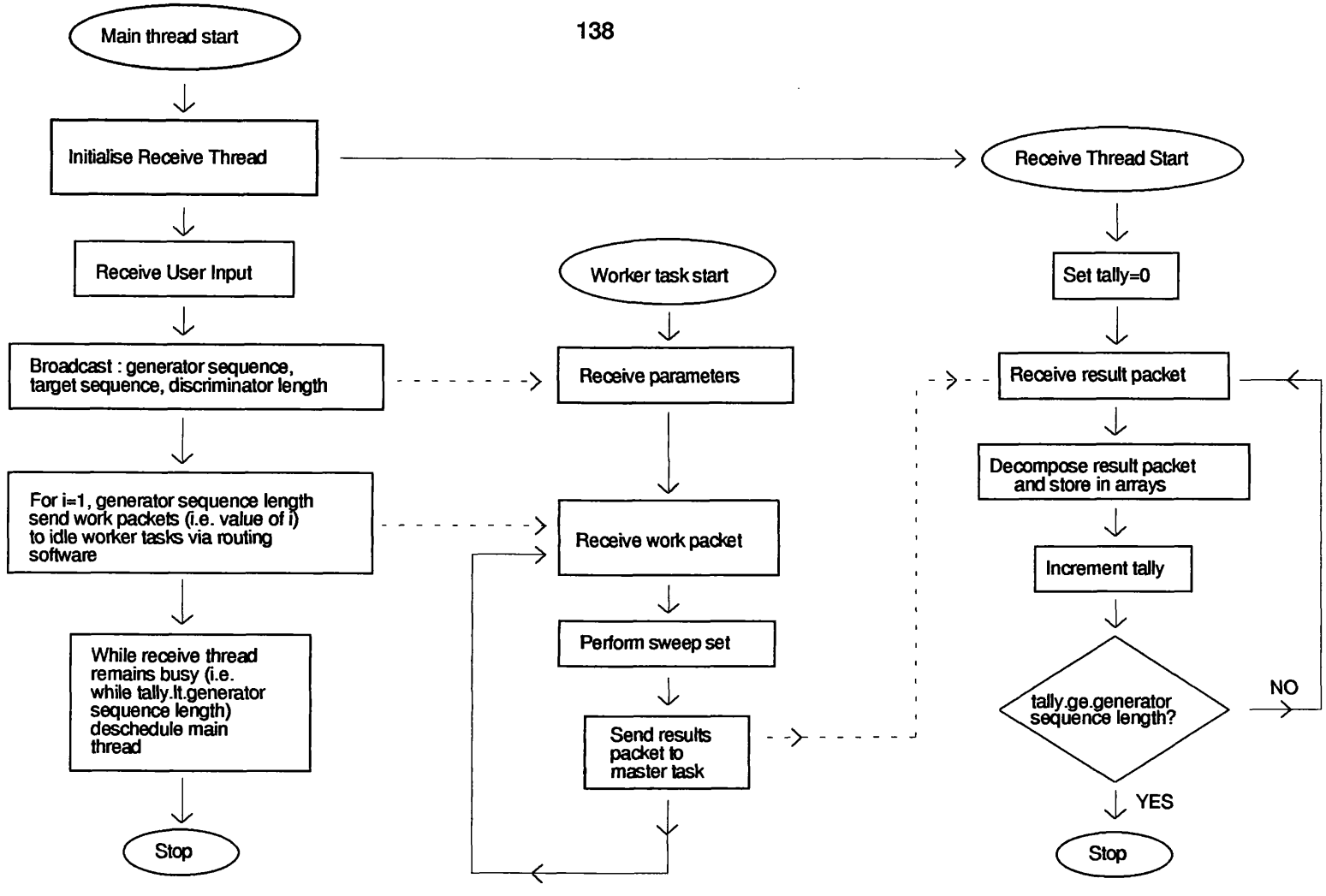


Figure 7.3.5.2; Overview of the design of the parallel DH_Scan program implemented as a Processor Farm. The dotted lines indicated message passing. The MAIN THREAD and the RECEIVE THREAD together constitute the master task, which resides on the root transputer. The worker task is automatically replicated on every transputer of the network at execution time. Optionally, a worker task can also be copied to the root transputer - this serves to mop up any spare capacity on the root transputer.

7.3.5.2.1 Algorithm - Master Task

The master task contains two threads: (i) the MAIN thread or main program; (ii) the RECEIVE thread. This is a slightly unusual format for the master task. Normally, the master task is split into three threads: main, send and receive. The main thread usually handles input and output only. In DH Scan the main thread also handles the forwarding of the work packets. The authors realised that the forwarding of work packets could be efficiently sandwiched between the input and output phases of the main thread without causing any problems.

The main thread notes start time by means of a "F77_Timer_Now()" function call and then initialises the receive thread by means of a function call, as follows:

```
      CALL F77_THREAD_START (RECEIVE, RECEIVE_WS,
     1        WS_SIZE*4,F77_THREAD_URGENT,0)
```

By means of a simple interactive session the main thread of the master task prompts the user (via the afserver) for the names of two homologous GPCRs. By default, the first protein sequence (the generator sequence, gs) is regarded as the sequence from which all the discriminators are generated and is read into a character array called discrim_file. The second protein sequence (the target sequence, ts) is read into an array called scan_file. The user is then prompted to enter the discriminator length.

Copies of each protein sequence are broadcast by the main thread to all of the worker tasks by means of two consecutive F77_Net_Broadcast_Send calls. A third F77_Net_Broadcast_Send function call forwards the discriminator length to all of the worker tasks. Work packets are generated and forwarded to idle worker tasks by means of a F77_NET_SEND call embedded in a simple DO loop, as follows:

```
      Do 100 i=1,generator_sequence_length
            data=i
            call F77_NET_SEND(4,data,true)
```

```
100 continue
```

On each cycle of the loop an integer value corresponding to a discriminator start position in the discrim_file is packed into a work packet and is automatically forwarded to an idle worker task by the routing software, if necessary passing through more than one node in order to reach one. For example, an interactive session in which human dopamine D2 (433 amino acids in length) and rat dopamine D3 were entered, would result in 433 work packets. Once all the work packets have been forwarded the main thread is descheduled by means of a DO WHILE loop, as follows:

```
do while (tally.le.generator_sequence_length)
 call F77_THREAD_DESCHEDULE
end do
```

The receive thread increments a tally (a globally declared integer value) after each result packet has been processed. When tally reaches a value equal to the total length of the generator sequence, main thread knows that receive thread has finished. Main thread then notes the finish time, writes ten output files to disc, closes each output file and informs the user that the DH Scan is complete by beeping.

7.3.5.2.2 Algorithm - Worker Task

The worker task consists of only one thread: PROGRAM align_w.f77. Firstly, each worker task waits for three broadcast messages to arrive. The first broadcast message provides the generator sequence (discrim_file); the second: the target sequence (scan_file); the third: a single integer informing the worker task the discriminator length (the number of amino acids in each discriminator remains constant throughout each DH Scan). The worker task then waits for the arrival of a work packet containing a single integer value corresponding to a unique start position in the generator sequence. From the start position a single no_gap discriminator is generated, followed by d one_gap_discriminators, d*d two_gap_discriminators, d one_del discriminators and d*d two_del_discriminators. ALL OTHER POTENTIAL START POSITIONS ARE IGNORED. Each discriminator

generated is moved along the whole length of the target sequence and scores computed on a pair-wise hit/miss basis and the highest score from the No_Gap_Sweep, One_Gap_Sweep, Two_Gap_Sweep, One_Del_Sweep and Two_Del_Sweep noted. Highest scores are packed into a results packet which is forwarded to the receive thread running in the master task on the root transputer. The worker task then waits to receive a new packet to allow it to carry out a new set of sweeps.

	Per Worker Task		$\sum$ Worker Tasks	
Type of Sweep	Number of Scans	Number of Pair-Wise Comparisons	Number of Scans	Number of Pair-Wise Comparisons
No Gap	1	$d.n_{ts}$	n_{gs}	$d.n_{gs}.n_{ts}$
One Gap	d	$d^2.n_{ts}$	$d.n_{gs}$	$d^2.n_{gs}.n_{ts}$
Two Gap	d^2	$d^3.n_{ts}$	$d^2.n_{gs}$	$d^3.n_{gs}.n_{ts}$
One Del	d	$d^2.n_{ts}$	$d.n_{gs}$	$d^2.n_{gs}.n_{ts}$
Two Del	d^2	$d^3.n_{ts}$	$d^2.n_{gs}$	$d^3.n_{gs}.n_{ts}$
Total	$2d^2+2d+1$	$n_{ts}(2d^3+2d^2+d)$	$n_{gs}(2d^2+2d+1)$	$n_{gs}.n_{ts}(2d^3+2d^2+d)$

Table 7.3.5.2.2 Breakdown of work loads for a generator sequence of length n_{gs} , target sequence of length n_{ts} and a discriminator length of d amino acids.

The advantage of the processor farm approach is immediately clear. Each amino acid of the generator sequence generates $2d^2+2d+1$ discriminators resulting in a total of $n_{gs}(2d^2+2d+1)$ sweeps of the scan_file and $n_{gs}n_{ts}(2d^3+2d^2+d)$ pair-wise comparisons (see above for table 7.3.5.2.2). On receipt of a work packet each worker task conducts $n_{ts}(2d^3+2d^2+d)$ pair wise comparisons. Since our approach is based on demand driven (task-farming) form of data parallelism, each transputer takes exactly the same amount of time to process a work packet. Hence, each transputer is used to an equal degree thus ensuring equitable load balancing. Scaling is achieved simply by adding more transputers to the network up to a maximum of n_{gs} transputers. Ignoring the few seconds for reading in the input files and writing output, running DH Scan on one transputer (that is, one worker task) would take: $n_{gs}*t_{wt}$ (where t_{wt} is the time taken, in minutes, by a busy worker task

to convert a newly arrived work packet into a results packet). n transputers would take: $n_{gs}*t_{wt}/n$ minutes and n_{gs} transputers would take just t_{wt} minutes.

We are of course ignoring the communication overhead, but this is minimised in our algorithm by keeping the t_{wt}/i_{wt} ratio very large (where i_{wt} is idle time on a single node due to communication overhead), i.e. $t_{wt} \sim 5$ mins (re: T800); $t_{wt} \sim 2$ mins (re: T805). We expect t_{wt} to be around 30 secs on a network of T9000 transputers. Since worker packets are buffered along the network (figure 7.3.1.2), an idle worker task does not have to wait for the main thread of the master task to send a fresh work packet. A queue of buffered work packets sits on top of each worker task. When a work task becomes idle, a work packet "drops" into the idle worker task allowing a new set of sweeps to be performed. In this way i_{wt} is minimised. Obviously, a massive parallel array of 400-500 transputers would mean that buffering and queuing of work packets would not occur. A massive parallel array of several hundred T9000 transputers would mean run times of around 40 seconds. Running computationally intensive applications at interactive speeds is the ultimate goal of parallelisation (Goodfellow *et al.*, 1990; Lomax *et al.*, 1991).

7.3.6 Implementation

7.3.6.1 Hardware

Our particular implementation of DH Scan uses an IBM compatible 386/387 MSDOS (version 5.0) PC with 1 Mbyte of RAM, a single five and a quarter inch disk drive, 80 Mbyte hard disk and a B004 plugin-card. The B004 board provided communication between the PC and a transputer-based machine designed. This transputer-based machine is made up of 32 T800 transputers with 4 Mbytes of RAM per transputer and is networked to several PC host machines each one of which is fitted with a B004 plugin-card. Hence, our implementation uses a parallel architecture best described as a distributed memory MIMD machine.

7.3.6.1.1 Heterogeneous Networks

DH Scan can be easily modified to run on a network made up of a mixture of T400 and T800/T805 transputers. The T400 transputer handles floating point operations differently from the T800/T805. The only part of the source code using floating point is a few lines in align_m.f77 which calculate the time elapsed in hours:minutes:seconds. Once these ten or so lines have been commented out DH Scan can be recompiled to run on T4s and T8s. If the timer functions are considered desirable, to run DH Scan on a mixture of T800 and T400 transputers requires a simple extension to the configuration file:

```
task t4master file=align_m4
task t8master file=align_m8
task t4worker file=align_w
task t8master file=align_m
```

Separate tasks must be compiled and linked for T4 and T8 transputers; the 3L Parallel FORTRAN compiler ensures that the right task images are loaded into the appropriate transputers in the network.

7.3.6.2 Software

The application was compiled, linked and configured using an MSDOS batch file, as follows:

```
REM Compile, link and configure the ALIGN application
t8f align_m /FL > fd_1
t8f align_w /FL > fd_2
linkt @align_m.lnk,align_m.b4
linkt @align_w.lnk,align_w.b4
REM configure `flood-filled' version of application
fconfig falign.cfg  falign.b4
```

The 3L supplied flood-filling configurer, FCONFIG was used to produce an executable file, as follows:

```
Task Master File=align_m Data=300k
Task Worker File=align_w Data=300K
```

The executable file automatically distributes copies of the worker task across an arbitrary network (hence the term "flood-filling"). The flood-filling configurer uses the 3L supplied frouter task which resides on each node of a flood-filled network. It is the frouter task, automatically copied to each node at execution time, which manages the flow of broadcast packets, work packets and results packets through the network.

The application is executed using another MSDOS batch file:

```
copy *.dat dat.bak
del *.dat
cls
afserver -:b falign.b4 -:o 1
```

The fast on-chip memory of the T800 is limited to 4KB. To ensure that the application is always allocated sufficient stack space the "-:o 1" switch is used. This switch (hyphen, colon, option letter "o", then a space, then the digit one) changes the way memory is allocated to give the application a very large amount of stack space. In our hardware implementation this is at least 4 Mbytes of RAM per transputer.

7.3.6.2.1 Tree Configurations

The processor farm approach can be implemented on three different configurations of transputers: the linear, binary and ternary tree structures. The linear tree network used in this research work offers the least efficient structure for implementation of the farm paradigm whilst a globally buffered ternary tree configuration offers the most efficient structure (Lomax *et al.*, 1991). The reason for this being that there are relatively more

transputers in the bottom layers of the ternary tree which have no results packets from other processors to deal with.

It is possible to alter the DH Scan code to take into account the type of tree network on which it is implemented. For example, the worker tasks could store results and forward them in a single results packet to the master task only when all of the work packets have been processed. For a network with the same number of worker tasks and work packets each transputer along a linear network would have n_{tn} - p_{tw} results packets to forward onto the next transputer. (Where n_{tn} is equal to the number of transputers in the network and p_{tw} is the position of each worker task along the network relative to the master task). Likewise, work packets could be formulated to take account of the number of transputers in the network - communication overhead would be minimised if the number of work packets equalled the number of transputers in a given network. Future versions of DH Scan will incorporate this and other features.

7.3.7 Runs

DH Scans were conducted between each member of the DA family of GPCRs (D_1 through to D_5) generating five sets of runs. Each set consisted of 4 separate runs. In the case of the D_1 set: D_1 with D_2, D_3, D_4 and D_5. To aid debugging and code checking the first two sets (the D_1 and D_2 sets) also included a *negative* control: the globular protein lysozyme from bacteriophage SF6 (Verma, 1986). Where time permitted, runs were also conducted between DA receptors and GPCRs belonging to other families, e.g. β_2-adrenergic receptor.

7.3.8 Prediction of Helix End Positions

DH Scans were examined. Where a distinct peak is observed, this is used to predict transmembrane helix start positions. For example, D_1/D_5 scan clearly reveals high discriminator homology - however, ptms I is difficult to distinguish from ptms II. In contrast, discriminator homology (as suggested by DH Stat value) with regard to D_2/D_5

scan is lower than for the D_1/D_5 scan - but ptms I is clearly visible as a separate entity in the scan. D_1/D_5 scans can be used to help predict the location of ptms 2 to 7 and D_2/D_5 to predict helix start position of ptms I (see table 7.3.8). An arbitrary number 25 is then simply added to the predicted start position to locate the putative end position to create a ptms of 26 residues in length. Engelman *et al.*, (1986) suggested that transmembrane helices are at least 20 amino acids in length in order to cross the lipid bilayer; a value of 26 allows the ptms to be tilted and has been used by other workers in the field, e.g. Dahl *et al.*, 1991b. Also, helices are likely to protrude into the intra and extracellular environment making a ptms of 26 residues in length a reasonable choice.

Table 7.3.8 Extract of discriminator homology from the D_2/D_5 DH Scan illustrating first peak corresponding to ptms I. There is a clear peak at positions 37/38. Adding 25 to 37 produces the putative helix end position to give: 37 - 62. This compares favourably with the SWISS-PROT prediction based on hydropathy analysis: 38 - 60.

Residue Position	DH Score
34	14
35	15
36	16
37	17
38	17
39	16
40	16
41	15
42	14
43	13

Where discriminatory homology peaks at a given value (x) and remains at that value a simple procedure was used to predict putative helix start position. For odd number of maximum DH values: the middle residue position within the constant set of discriminator values was chosen as the helix start position and 25 was added to such values to predict helix end position. For even number of maximum values: n is added to the position corresponding to first occurrence of x to yield the putative helix start position; where : $n = y/2$; y = length of sequence with DH scores of x Armed with the location of ptms I (in D_2) and ptms II to VII (in D_1) manual alignment was used to locate the transmembrane helices in each subtype DA receptor.

7.4 Results and Discussion

It is clear from the DH Scans (figures 7.4.1 to 7.4.5; pages: 151 to 155) that two sub-families exist within the dopamine group of GPCRs. D_1 and D_5 form one sub-group and D_2 and D_3 and D_4 form the second subfamily which is in agreement with the suggested classification scheme of Sibley and Monsama (1992). However, D_4 appears to differ considerably from D_2 and D_3. One possible interpretation for the lack of strong discriminator homology (also reflected in the DH Stat scores - see table 7.4.1 on page 156) is that D_4 is not a true member of the D_2 sub-family of DA GPCRs. D_4 might possibly be the first cloned member of a third sub-family of DA receptors. Using the Sibley and Monsma (1992) preferred classification scheme, D_4 would be known as D_{3A} rather than D_{2C}. Though a systematic study was not carried out, a curious feature of the DH Scans and the DH Stat scores is that the DA receptors sometimes exhibit stronger discriminator homology with GPCRs from other families. For example, D_2 has stronger discriminator homology with A2A-adrenergic receptor than D_1 (and almost by default: D_5). Side effects from various drug treatments which target the DA receptors are well known so this observation is not surprising.

From the perspective of sequence analysis there are several methods which could be used to distinguish members of this group. The method preferred by Sibley and Monsama (1992) involves conducting transmembrane homology studies. They suggest that if the transmembrane homology between two or more dopamine sequences is greater than 50%, then these molecules belong to the same subgroup within the dopamine family. Unfortunately, this method relies on the accuracy of the putative locations of the transmembrane helices. Fasman and Gilbert (1990) point out that current techniques based on various forms of hydropathy analysis to predict the locations of transmembrane domains are not always reliable. DH Scan provides a simple alternative method to predicting the location of transmembrane helices (see table 7.4.2 on page 157). Though DH Scan by default incorporates the ideas of Sibley and Monsama (1992) with regard to exploiting transmembrane homology to correctly classify GPCRs it was never-the-less coded completely independently. Interestingly, DH Scan clearly indicates that members of the D_1

subfamily should also exhibit high homology in the terminal region of the carboxy tail (figures 7.4.1 and 7.4.5)[1]. A feature not found in the D_2 subfamily of DA receptors.

Given that DH Scans so clearly illustrate homology clusters or peaks corresponding to regions of high transmembrane homology. It follows that DH Scans could be used to aid the construction of chimeric receptor proteins, in which regions from two GPCRs are linked and the pharmacological phenotype of the resulting hybrid receptor protein determined. In this way the molecular basis for DA receptor subtype specificity could be studied. For example, it is particularly noticeable that discriminator homology is low with regard to ptms IV and V between members of the D_1 and D_2 sub-families. A chimeric receptor in which the ptms IV and then ptms V is copied from say D_2 or D_3 to D_1 or D_5 and subsequent phenotyping could confirm the origin of subtype specificity within the DA family of GPCRs. Such data would be invaluable to aid medicinal chemists in the rational design of selective drugs. Chimeric proteins have been constructed between for example: β_1/β_2-adrenergic receptors (Frielle *et al.*, 1988) and more recently between D_1/D_2 which produced a chimeric dopamine GPCR which mediates a D_1 response to a D_2 selective agonist (MacKenzie *et al.*, 1993).

Given that side-effects are often a problem with new drug treatments - DH Scans applied between DA receptors and all other human GPCRs cloned to date and published in the form of a simple source book would certainly help to warn medicinal chemists about the possibility of interaction of their new compounds with other GPCRs which exhibit high discriminator homology. DH Scans are easy to interpret. This is a very important point given that medicinal chemists (and many biologists - Gribskov and Devereux, 1991) typically receive no formal training in computerised sequence analysis. Interpreting matrix dot plots is not for them. They want a scan that is easy to interpret - preferably one that looks like an NMR spectra! Given the ease with which standard graphics libraries can be used in displaying data in a highly visual way - future versions of the algorithm could make use of colour. Certainly, a source book containing hundreds of DH Scans would be a very useful tool to aid the medicinal chemists in the rational design of selective drugs.

[1] Sibley and Monsama had not observed that members of the D_1 subfamily also exhibit high homology in the terminal region of the carboxy tail.

One often quoted method is based on the Needleman and Wunsch (1970) algorithm. For example, carrying out a Needleman and Wunsch sequence alignment between members of the dopamine family will provide percentage identity values (results not shown). Using GAP (Devereux *et al.*, 1984) to carry out sequence randomizations (reviewd by Gribskov and Devereux, 1991) provides a basis to clearly distinguish between different members of the dopamine family. The method can be performed on a VAX taking a few minutes for each GAP/RAN=n command to run on a multi-user VAX 8650 under medium load - a farm-processor parallel implementation of this method would actually take just a few seconds allowing the complete table to be quickly built up. However, this method fails to reveal the varying degrees of sequence conservation along each protein sequence. In addition, the algorithm adds up the highest discriminator scores (equivalent to integrating the area under the DH Scan curve). This cumulative score (DH Stat - see table 7.4.1) also provides a simple basis to quantitatively differentiate between closely related proteins.

Less homologous GPCRs are likely to become more amenable to our algorithm when favoured amino acid substitutions are considered. For example, Chung *et al.* (1987) who found that homology between human brain beta-receptor and pig brain muscarinic receptor increased from 31% to 46% when using favoured amino acid substitutions as defined by Dayhoff *et al.* (1978). Employing similarity matrices based on for example physio-chemical properties will allow future versions of DH Scan to clearly display degrees of residue character conservation, as distinct from sequence conservation. Attwood *et al.* (1991) have already pioneered a similar approach using database pattern-matching discriminators. Our approach however will generate all possible pattern-matching discriminators (with insertions and deletions) from a single generator sequence - something really only possible with the computational power offered by massive parallel processing. In addition, we intend to support a particularly flexible sequence pattern matching syntax (Jones, 1992) to aid in the detection of regions of high functionality.

7.4.1 Current And Future Trends

This work was conducted in early 1992 and there have been considerable developments on the supercomputing field since then. The promised T9000 transputer (the replacement to the T800 series) has been shipped in quantities but only after considerable delays and only offers approximately one order of magnitude improvement in performance over the T800 Series. The saying "technology waits for no one" is particularly valid in this area. Consequently, parallel technology from Intel (their iPSC/860 technology which offers in excess of 50 Mflops (Mega Flop) per node) and the 50 MFlop C40 Texas Instruments chip has rather pushed the T9000 Series into the second league in terms of performance. Companies such as Parsytec offer GFlop machines *off the shelf* based on Intel iPSC/860 parallel technology. In addition, Cray have 11 orders for their new TD3 parallel supercomputer.

Given that machines are now available offering TFlop (Teraflop) performance the use of the T9000 series transputer from Inmos in Grand Challenge projects looks increasingly vulnerable. In addition, the release of 64 bit RISC (Reduced Instruction Set) chips from DIGITAL™ (installed in the Alpha DEC workstations) which run standard FORTRAN and C compiled code without the additional cost of communication overhead renders current T800 transputer based architectures obsolete and T9000 based transparallel hardware looks increasingly vulnerable to competition from later versions of the Alpha DEC chips. In short, transputers of any type are no longer the competitive choice in terms of number crunching capability though they have a role in embedded systems, e.g. a T400 Series transputer is used in each Intel iPSC/860 node to control message passing, though it is does not directly contribute to the number crunching capability of the iPSC/860. Likewise, the PowerTRAM from Parsytec of Germany combines the performance of the *state of the art* RISC processor Power PC™ 601 (80mhz, 160MIPS/80MFLOPS peak performance) with transputer communication capability of the IMS (Inmos) T425 communication processor (25mhz, 4 serial links each with 20Mbit/s transfer rate - Parsytec Data Spec Sheet for the PowerTRAM 601/80, 1994).

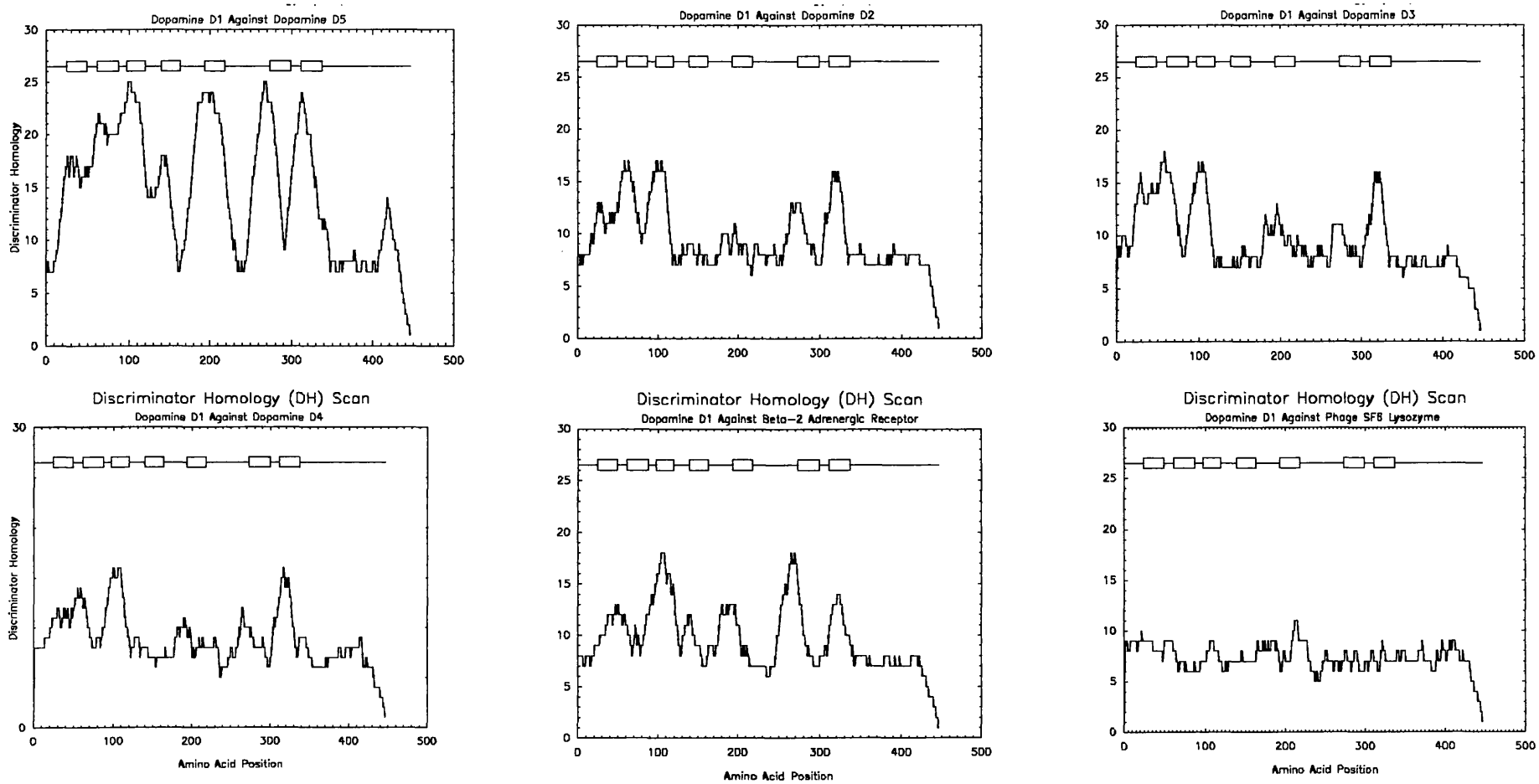


Figure 7.4.1; DH Scans of human D_1 receptor sequence against each member of the dopamine family of G protein coupled receptors (GPCRs), β_2 adrenergic receptor (Chung *et al.*, 1987) and lysozyme of phage SF6 (Verma, 1986) which provides a *negative* control. Discriminator length was 25 residues in each scan. Positions of putative transmembrane segments (ptms) of D_1 based on hydropathicity (conversely hydrophobicity) analysis as given in SWISS-PROT are represented rectangular box shapes above each scan. Use of SWISS-PROT transmembrane predictions based on hydropathy analysis and displaying these on each DH Scan avoids experimenter bias. It is clear that D_1 is closely related to D_5 - transmembrane homology exists between each ptms. Of particular note is the clear peak corresponding to ptms VII. Ptms VII is difficult to *see* by means of hydropathy plots due to its amphiphilic nature (chapter 5). Also, it is clear that transmembrane homology is less for ptms IV and to a slightly lesser extent ptms V. This suggests that ptms I, II, III, VI and VII each contribute to providing a ligand binding pocket. DH_Stat scores are: 6585 (D_1/D_5), 4247 (D_1/D_2), 4261 (D_1/D_3), 3930 (D_1/D_4), 4313 (D_1/β_2) and 3291 (D_1/SF6) and is in agreement with the consensus that D_1 and D_5 belong to the same sub-family of GPCRs. β_2 receptor exhibits closer discriminator homology with D_1 than D_3 and D_4. This suggests that ligands designed to bind to D_1 may also bind to β_2 (and possibly other GPCRs). In the DH Scan of D_1 against D_5 their is an "eighth" peak nearly one hundred residues downstream of ptms VII in the terminal region of the intracellular carboxy tail of the receptor. The aligned section is show below; capital X indicates residue conservation:

```
D1 SVI_LDYDTDVSLEKIQPITQNGQH  421-444
D5 SVWELDCEGEISLDKITPFTPNGFH  453-477
     XX__XX_X_XXXXXXX_X_X_XX_X
```

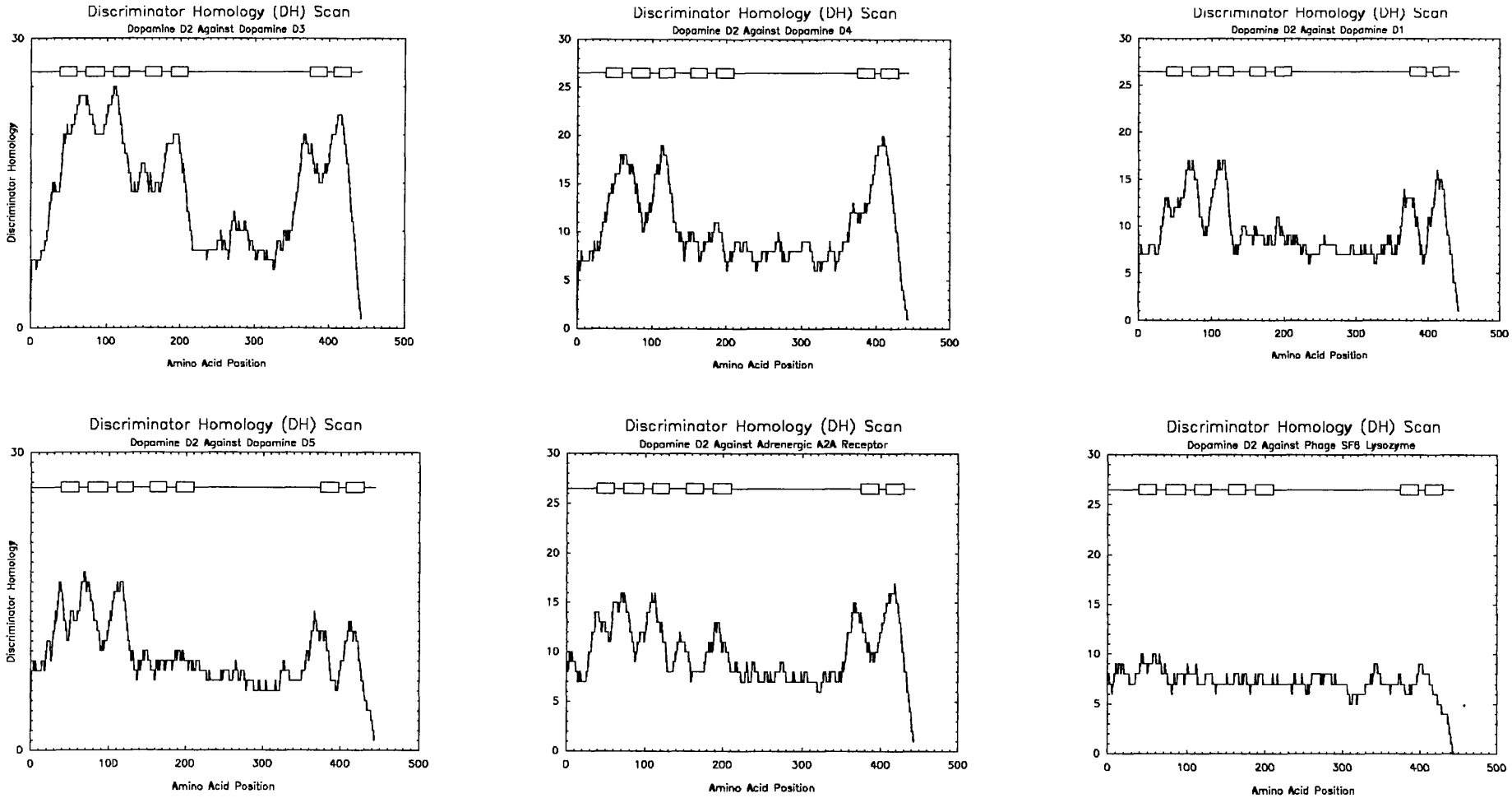
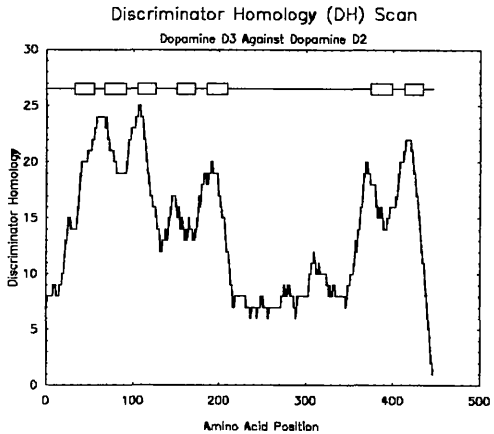
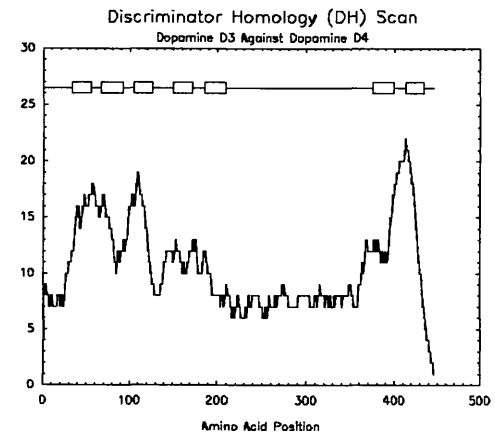
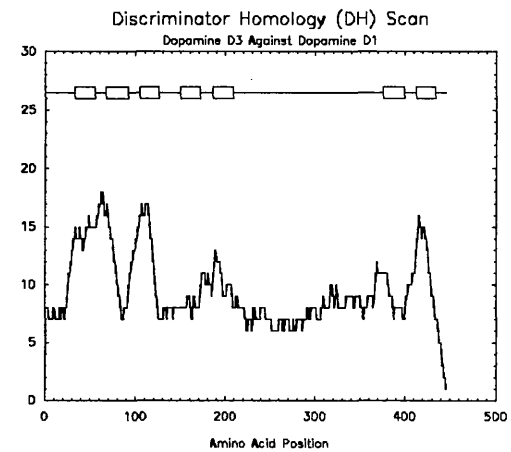


Figure 7.4.2 DH Scans of human D_2 receptor sequence against each member of the dopamine family of G protein coupled receptors (GPCRs), A2A adrenergic receptor (Fraser *et al.*, 1989) and lysozyme of phage SF6 which once more provides a *negative* control for comparison. Discriminator length was 25 residues in each scan. Positions of putative transmembrane segments (ptms) of D_2 based on hydropathicity (conversely hydrophobicity) analysis as given in SWISS-PROT are represented rectangular box shapes above each scan. It is clear that D_2 is closely related to D_3 - transmembrane homology exists between each ptms. Of particular note is the clear peak corresponding to ptms VII. Ptms VII is difficult to *see* by means of hydropathy plots due to its amphiphilic nature (chapter 5). Also, it is clear that transmembrane homology drops off with regard to ptms IV and V. This strongly suggests that ptms I, II, III, VI and VII each contribute to providing a ligand binding domain. Comparison with the D_1 set of DH Scans clearly reveals differences in the overall topology; the D_2 dopamine (DA) receptor has a long sequence of residues between ptms V and VI and a much shorter tail sequence than the D_1 DA receptor. In addition, the "eighth" peak clearly noticable in the D_1/D_5 (and by default in the D_5/D_1 scan) is missing in the D_2/D_1 scan. DH_Stat scores are: 4314 (D_2/D_1), 6327 (D_2/D_3), 4617 (D_2/D_4), 4330 (D_2/D_5), 4414 (D_2/A2A), 3204 (D_2/SF6). The DH_Stat score for D_2/A2A scan is higher than either D_2/D_1 or D_2/D_5 scans. This strongly suggests that ligands designed to have high affinity for D_2 (and almost by default: D_3) but with low affinity for D_1 (and by default: D_5) might unfortunately also have high affinity for other receptors such as A2A-adrenergic receptor. This possibility is born out in practise







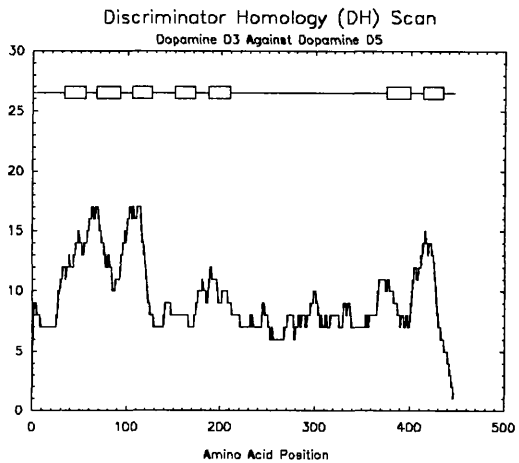
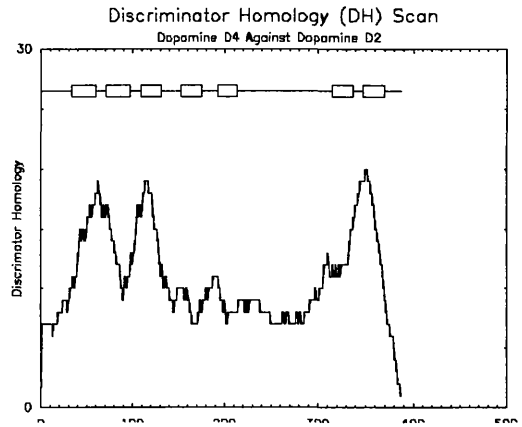
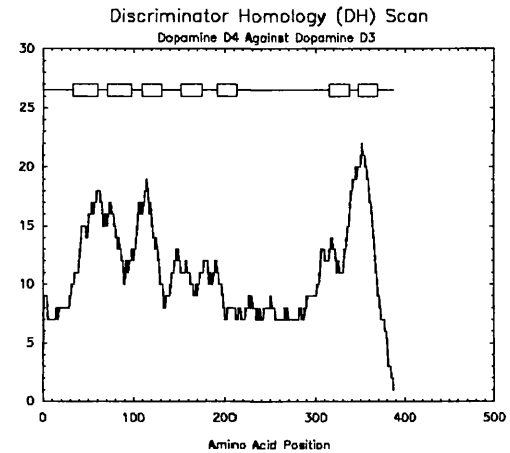
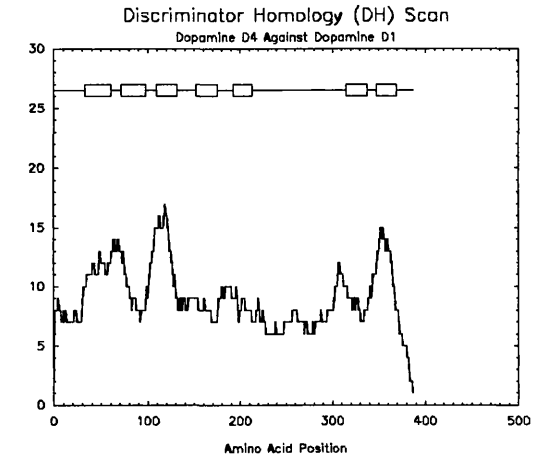


Figure 7.4.3 DH Scans of human D_3 receptor sequence against each member of the dopamine family of G protein coupled receptors (GPCRs). Discriminator length was 25 residues in each scan. Positions of putative transmembrane segments (ptms) of D_3 based on hydropathicity (conversely hydrophobicity) analysis as given in SWISS-PROT are represented rectangular box shapes above each scan. It is clear that D_3 is closely related to D_2 - transmembrane homology exists between each ptms. The D_3/D_1 scan is very similar to the D_3/D_5 scan in agreement with the consensus that D_5 is a member of the D_1 subfamily of dopamine GPCRs.







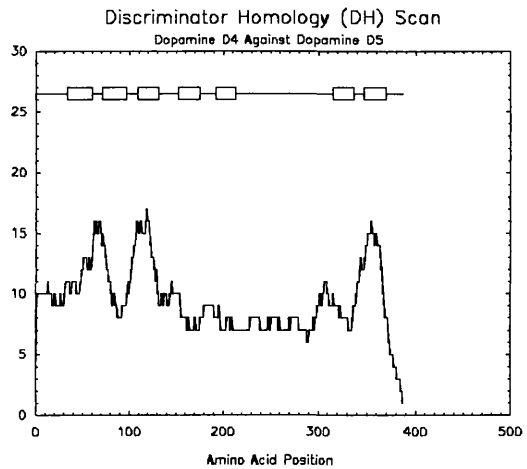
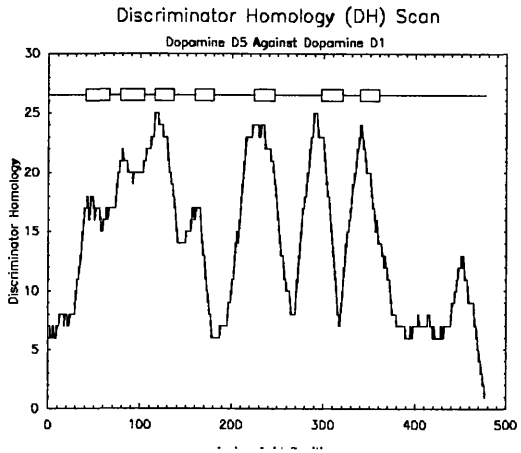
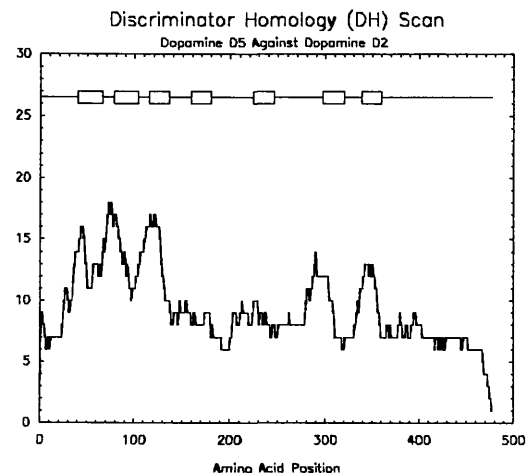
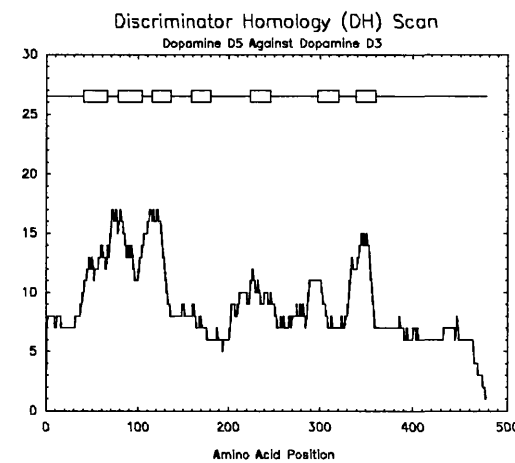


Figure 7.4.4 DH Scans of human D_4 receptor sequence against each member of the dopamine family of G protein coupled receptors (GPCRs). Discriminator length was 25 residues in each scan. Positions of putative transmembrane segments (ptms) of D_3 based on hydropathicity (conversely hydrophobicity) analysis as given in SWISS-PROT are represented rectangular box shapes above each scan. With regard to the D_4/D_1 and D_4/D_5 scans it is clear that transmembrane homology does not exist between each ptms in agreement with the consensus that D_4 is a member of the D_2 subfamily of dopamine GPCRs.







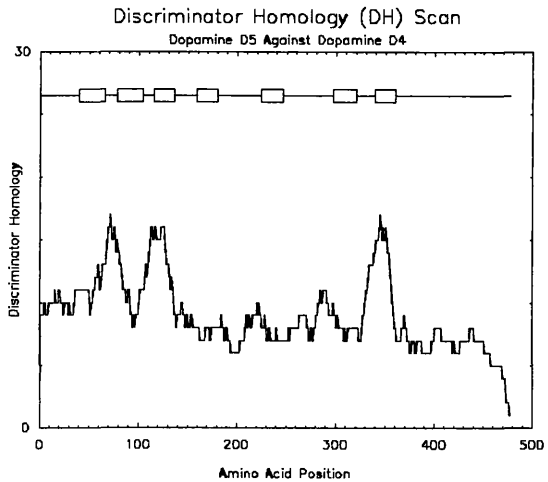


Figure 7.4.5 DH Scans of human D_5 receptor sequence against each member of the dopamine family of G protein coupled receptors (GPCRs). Discriminator length was 25 residues in each scan. Positions of putative transmembrane segments (ptms) of D_5 based on hydropathicity (conversely hydrophobicity) analysis as given in SWISS-PROT are represented rectangular box shapes above each scan. It is clear that D_5 is closely related to D_1 - transmembrane homology exists between each ptms. The scans are in agreement with the perceived characteristic topology of the D_1 subfamily of dopamine (DA) GPCRs: a short third intracellular loop between ptms V and VI and an extended carboxy tail sequence. In addition, it is noticeable that their is high homology in the terminal region of the carboxy tail of D_1 and D_5 GPCRs.

	D_1	D_2	D_3	D_4	D_5
D_1	X	4247	4261	3930	**6585**
D_2	4314	X	**6327**	4617	4330
D_3	4396	**6347**	X	4843	4263
D_4	3592	4292	**4405**	X	3786
D_5	**6886**	4558	4404	4385	X

Table 7.4.1 Table of DH_Stats. These represent the total area under each DH Scan - i.e. total number of hits based on pairwise identity. Discriminator size = 25 residues for each DH Scan. The table reads down column one and across. Bold is used to draw attention to highest scores. D_1, D_2, D_3 and D_5 receptors have particularly strong discriminator homology with one other dopamine (DA) receptor (i.e. D_1 with D_5, D_2 with D_3 and visa-versa). In contrast D_4 lacks a distinctly high score even though it is considered to be a member of the D_2 subfamily of DA receptors suggesting that divergent evolution is in play. The matrix is not symmetric since primary structures are not equal in length and the discriminators are generated from only one sequence (not both sequences) during each DH Scan. Consequently, DH Scan of e.g. D_1 against D_5 has a lower score than the DH Scan of D_5 against D_1: 6585 and 6886 respectively. [D_1 is 446 residues in length and D_5 is 477 residues in length - hence the y axis of the coresponding plots for each DH Scan are: 446 residue positions and 477 residue positions respectively].

PTMS	D_1		D_2		D_3		D_4		D_5	
	DH	SW	DH	SW	DH	SW	DH	SW	DH	SW
I	25-50	24-49	37-62	38-60	32-57	33-55	37-62	33-60	43-68	40-66
II	63-88	61-87	73-98	72-97	68-93	67-92	73-98	71-97	80-105	78-104
III	100-125	97-119	111-136	109-130	107-132	105-126	112-137	109-131	117-142	115-136
IV	141-166	139-163	153-178	152-174	151-176	150-172	153-178	152-175	158-183	159-180
V	194-219	193-218	189-214	187-210	188-213	186-209	191-216	192-213	225-250	224-246
VI	268-293	273-299	369-394	374-397	371-396	375-399	309-334	315-337	292-317	297-320
VII	312-337	311-337	408-433	406-429	410-435	413-434	348-373	347-369	340-365	339-360

Table 7.4.2 Prediction of putative transmembrane segments (ptms - otherwise referred to as spanners or transmembrane helices) in dopamine (DA) receptors: D_1 to D_5. Discriminator Homology Scan (DH) was used in the following way: ptms I was determined from DH Scan of D_2 against D_5. All other ptms I's (i.e. in D_1, D_3, D_4 and D_5) were obtained by manual alignment with the predicted ptms I of D_2. DH Scan of D_1 against D_5 were used to predict location of ptms II to VII (in D_1) and by simple manual alignment the remaining ptms in the rest of the DA receptors were located (more details given in main text). Predictions of ptms by the method of hydropathy analysis as recorded in the SWISS-PROT (SW) database are included for comparison. Hence, only two DH Scans were required (D_2/D_5 and D_1/D_5) to predict all DA receptor transmembrane spanners. Since DA G protein coupled receptors (GPCRs) exhibit significant sequence homology with for example beta-adrenergic receptors (see chapter 5) - the ptms of these GPCRs could be predicted from just these two DH Scans as well. It is clear that DH Scan provides an alternative method for predicting transmembrane helices in DA receptors.

7.5 Conclusion

Smith (1990) considered parallel processing from three perspectives: "load balancing", "communication" and "scaling". Load balancing, as the term implies, means that the program executing on a transputer-based parallel computer should, as far as possible, use each transputer simultaneously and to an equal degree; communication should be kept to the minimum to avoid costly overheads; scaling should be easy to achieve. It is clear that on all three accounts, the novel sequence alignment algorithm (DH Scan) scores highly. That this is so, merely reflects the ease with which a processor farm can be implemented on a large transputer network.

This is has been an unusual piece of work - if only because parallel computing is rarely used in biology. Consequently, some effort has been applied to help educate the reader/biologist in the business of parallel computing as implemented using Inmos transputer technology. DH Scan was written to specifically use this technology and to apply it to DA GPCRs. DH Scan is clearly a useful tool in helping to discern the seventh putative transmembrane helix which is not seen in hydropathy plots and to provide an independent verification of the likely starting points of transmembrane spanners. It also provides an unambiguous view of the positions of clustered sequence identities thereby showing that D_2 is clearly closely related to D_3 and less closely related to any other member of the DA family of GPCRs. DH Scan also clearly demonstrates that D_1 is closely related to D_5. Consequently, 3D models of D_2 and D_1 would provide useful structural templates to model D_3 and D_5 respectively. In addition, DH Scan can be used to suggest chimeric genetic analysis experiments.

8. Fourier Sequence Analysis of Catechol Amine Binding G Protein Coupled Receptors - Implications For The Three Dimensional Structure And Binding of Agonists

8.1 Summary

Fourier analysis of a multiple-sequence alignment suggests that the consensus view that only those residues facing the protein interior are conserved is not correct. In particular, transmembrane helices II and III do not exhibit residue conservancy characteristic of an amphipathic helix. It is proposed that these two helices undergo a form of helix interface shear to assist agonist binding to a Asp residue on helix II. The role of Fourier analysis in establishing the likely orientation of transmembrane helices in the catechol amine binding G protein coupled receptors (which includes the dopamine family of receptors) is discussed.

8.2 Introduction

Characterisation of residues in contact with the lipid bilayer is important for structural analysis of membrane proteins (Yeates *et al.*, 1987; Rees *et al.*, 1989a, 1989b). These workers identified residues in contact with the lipid bilayer by means of Fourier Analysis of multiple sequence files (MSF). Komiya *et al.* (1988) stated that: "This method may prove useful for modelling the three-dimensional structures of membrane proteins".

G protein coupled receptors (GPCRs) are integral membrane proteins and catechol amine binding GPCRs (the dopamine family) are particularly important in medicine. It makes perfect sense to apply the techniques used by the earlier workers to characterise the transmembrane region of photosynthetic reaction centre (RC) to also characterise the relative orientation of the transmembrane helices of dopamine (DA) family of GPCRs. Such information provides the first step in obtaining a 3D model of the transmembrane region of GPCRs (Donnelly *et al.*, 1989). Indeed, Fong *et al.* (1993) have applied Fourier Analysis to the transmembrane region of neurokinin-1 receptor (NK-1R) to locate residues

likely to be facing the interior of the protein. The general consensus being that surface residues (i.e. those facing the surrounding lipid) are less well conserved in homologous membrane proteins than the buried, interior residues (for example: Rees *et al.*, 1989b).

Eisenberg *et al.* (1984) have used the hydrophobic moment to detect periodicity in protein hydrophobicity. A periodicity in the hydrophobicity of 3.6 residues is characteristic of an amphiphilic α-helix. Amphiphilic β-sheets were found to have a periodicity of 2.3 residues, 3_{10} helices displayed a periodicity of about 2.5 residues. It was concluded that many protein sequences tend to form segments of maximum amphiphilicity . The hydrophobic moment can be calculated for any segment of a known primary structure. However, it is also conceivable to calculate conservancy moment profiles using multiple sequence files containing homologous proteins and thereby detect amphipathic[1] secondary structures.

The Fourier analysis package: Peppi! has been coded to produce the following:

- variability plots (Donnelly *et al.*, 1989).
- conservancy moment profile plot (modification of work performed by Eisenberg *et al.*, 1984).
- Fourier transform power spectrum P(ω) plot (Komiya *et al.*, 1988).
- calculate ψ (the average value of P(ω) in the α-helical range ($90^0 < \omega < 120^0$); Komiya *et al.*, 1988.
- a sliding window version of ψ used to predict the presence of amphipathic α-helices (Rees *et al.*, 1989b).
- helical wheel plots of the transmembrane region of selected GPCRs showing the probable orientation of transmembrane α-helices (Baldwin, 1993).
- vertical plots of transmembrane α-helices showing sequence variability at each residue position (used by Donnelly *et al.*, 1989).

[1] The term *amphipathic* is used in this work to describe protein segments which exhibit residue conservancy characteristic of a secondary structure with conserved residues on one side (β-sheet) or face (α-helix).

8.3 Methods and Materials

Two versions of Peppi! have been coded (for source code see: appendix 3). One version runs on transputer based kit and the other runs on Alpha DEC AXP 3000 series workstations. The Inmos transputer version is written in 3L Parallel FORTRAN (version 2.1.2) and the Alpha DEC Open VMS version is very similar but designed to compile using the standard VAX/VMS FORTRAN (version 6.1) compiler using DEC GKS (Graphic Kernal System; version 5.2) calls for producing all output to screen and printers. DEC GKS compiles with the international standard ISO 8805(E)-1985 and is an upwardly compatible extension to ISO GKS Standard 7942-1985. Hence this version is easily ported to any system using GKS which meets the stated ISO international standards, e.g. ULTRIX systems (DEC GKS Users Guide - June, 1992) and Alpha DEC AXP workstations running the OSF/1 implementation of the UNIX operating system.

The transputer version is not portable as it uses non-standard graphics library calls developed in-house by Noel Ruddock (Laboratory of Molecular Modelling, Glasgow University) and uses drivers for non-standard kit. The GKS version allows the user to send output to: black and white postscript, colour postscript and HPGL (Hewlett Packard Graphics Language) colour pen plotters as well as screen and/or disc; GKS supports a wide range of output devices and so the GKS version of Peppi! can be easily appended to permit output to these devices as well. The 3L version only writes output to the screen and/or disc; output to other devices is severely limited as the current version of the graphics library does not support a range of output devices.

The user edits a file (use.dat - see appendix 3) to select type of processing or output required. This file is read by the executable at run time. This allows the program to act on the users instructions immediately at run time without requiring the user to respond to layers of menus. Peppi! requires a multiple sequence file (MSF) in the popular GCG (Genetics Computer Group; Devereux *et al.*, 1984) format - typically produced using the GCG PILEUP command. For the MSF generated for this study see appendix 1.

8.3.1 Variability Profile Plots

The suggested transmembrane assignments of Baldwin (1993) were used throughout the Fourier analysis study. The residue variance (V_j) at each residue position is calculated from a family of aligned GPCRs - see figure 8.2.1. The V_j element of this profile is defined by the number of different types of amino acid residues that are observed at a given position j. More variable regions are likely to be exposed to the surrounding lipid (for example: Rees *et al.*, 1989b).

8.3.2 The Fourier Transform Power Spectrum P(ϖ)

To search for periodicity's in a predicted secondary structure, the Fourier transform power spectra, P(ω), is calculated:

$$P(\varpi) = \left[\sum_{j=1}^{N}\left(V_j - \bar{V}_j\right)cos(j\varpi)\right]^2 + \left[\sum_{j=1}^{N}\left(V_j - \bar{V}_j\right)sin(j\varpi)\right]^2 \qquad \textbf{Equation 1}$$

Where N is the number of residues in the sequence; ω is the angular rotation angle between residues around a helical axis (it equals 100^0 for a regular α-helix); ∇_j is the mean value of V_j for the sequence.

8.3.3 Measure of the α-Helical Character of the P(ω) Plot - ψ

The α-Helical Character of the P(ω) plot for a particular segment is described by the parameter ψ (Komiya, *et al.*, 1988):

$$\psi = \left[(1/30)\int_{90}^{120} P(\varpi)d\varpi\right] \Big/ \left[(1/180)\int_{0}^{180} P(\varpi)d\varpi\right] \qquad \textbf{Equation 2}$$

The y-axis of graphical output was automatically scaled using the highest peak as a guide. To avoid poor scaling, the first dominant peak was removed by a filter to give ψ_w :

$$\psi_w = \psi - \rho \qquad \textbf{Equation 3}$$

Where ρ is the area under the first peak of the Fourier transform spectrum. So ψ_w approximates to:

$$\psi_w \approx \left[(1/30) \int_{90}^{120} P(\varpi)\, d\varpi \right] \Big/ \left[(1/160) \int_{20}^{180} P(\varpi)\, d\varpi \right] \qquad \textbf{Equation 4}$$

For the purposes of the ψ_w calculation only the first peak is removed.

8.3.4 Sliding Window Version of ψ_w

ψ_w values were calculated over the whole length of the MSF file. The approximate version for ψ_w was used throughout (i.e. the area under the curve from 0^0 to 20^0) was ignored. The computation of the sliding version of ψ_w is quite demanding on the CPU. Since the Alpha chips are rated at something like 50 Mflops the entire sweep of the MSF file is performed in less than a minute. However, the Inmos T800 transputer is rated at around 1.5 MFlops and the sweep can take up to 40 minutes. The sliding version of ψ_w certainly amenable to farm processing form of implicit parallelism - see chapter 7. However, the transputer version of the code is not likely to be used by anyone else given its lack of portability and is unlikely to be updated. The GKS version of the code will be subject to future improvements only.

8.3.5 Conservancy Moment Plot

The conservancy moment can be calculated using the Fourier transform method used by Eisenberg *et al.* (1984) who used the method to calculate hydrophobic moments and to plot hydrophobic moment profiles. The conservancy moment, C(ω) can be calculated using the formula:

$$C(\varpi) = \left\{ \left[\sum_{j=1}^{N} c_j \sin(j\varpi) \right]^2 + \left[\sum_{j=1}^{N} c_j \cos(j\varpi) \right]^2 \right\}^{1/2} \qquad \textbf{Equation 5}$$

Where c_j is the residue conservancy (i.e. number of sequences / V_j).

The conservation moment is the modulus of the Fourier transform of the conservancy function:

$$C(\varpi) = \left| \sum_{j=1}^{N} c_j \varepsilon^{ij\varpi} \right| \qquad \textbf{Equation 6}$$

However, Peppi! uses equation 5 to produce conservancy moment profiles for a given length of MSF file.

8.3.6 Helical Wheel Plots Of The Transmembrane Region

Peppi! plots helical wheel schematics of the transmembrane region are based on Fourier analysis of the MSF dominated by catechol amine binding GPCR sequences - including each member of the dopamine family (D_1 .... D_5). The arrangement of the helices in the plot follows the probable arrangement of helices in GPCRs (Baldwin, 1993) rather than bacteriorhodopsin (bRh; Henderson *et al.*, 1990) which does not have any significant homology with any GPCR (see chapter 5). Baldwin based her findings on the experimentally derived low resolution two dimensional structure of rhodopsin (rH) obtained by Schertler *et al.*, (1993) rH is a genuine GPCR and has significant homology with the dopamine family of GPCRs (chapter 5).

Periodicity Analysis Algorithm: PEPPI!
VAX GKS$ version v1.0

Vj Scores for helix number: 1
Multiple sequence file (msf) name: gosh1.dat
Start and end positions (re: msf): 75, 100

```
  D2 Human    N Y Y A T L L T L L I A V I V F G N V L V C M A V S
    D3 Rat    A Y Y A L S Y C A L I L A I I F G N G L V C A A V L
  D4 Human    A A A L V G G V L L I G A V L A G N S L V C V S V A
 A2a Human    L T L V C L A G L L M L L T V F G N V L V I I A V F
 A2c Human    A G L A A V V G F L I V F T V V G N V L V V I A V L
 A2b Human    A A I A A A I T F L I L F T I F G N A L V I L A V L
5hta Human    V I T S L L L G T L I F C A V L G N A C V V A A I A
  D1 Human    I L T A C F L S L L I L S T L L G N T L V C A A V I
  D5 Human    V V T A C L L T L L I I W T L L G N V L V C A A I V
  B1 Human    . M . G L L M A L I V L L I V A G N V L V I V A I A
  B2 Human    . M . G I V M S L I V L A I V F G N V L V I T A I A
  B3 Human    . L A G A L L A L A V L A T V G G N L L V I V A I A
 5ht2 Rat    N W S A L L T T V V I I L T I A G N I L V I M A V S

 Vj =        8 9 8 5 6 6 8 6 5 4 3 6 7 4 3 5 1 1 7 2 1 3 6 2 2 6
```

Figure 8.2.1 Calculation of the variability profile based on transmembrane helix I (begin and end positions are those of Baldwin, 1993). V_j is the variability index calculated from the multiple sequence alignment file (MSF). V_j is simply the number of different residues occupying position j in the MSF segment. Each gap (represented by a dot) increases V_j score by 1 unit.

8.4 Results and Discussion

8.4.1 Fourier transform methods

Fourier transform methods provide a quantitative approach for characterising the periodicity of conserved and variable residues in a family of aligned sequences. It is clear from the variability plots, conservancy moment profile plots, table of ψ values, Fourier transform power spectra and psi analysis that helices I, IV, V and VII are amphipathic (figures 8.4.1.1 to 8.4.1.4 respectively) that helices II, III and VI are clearly not amphipathic (also see table 8.4.1 below).

TRANSMEMBRANE HELIX	ψ
I	2.4
II	1.4
III	1.1
IV	2.1
V	2.3
VI	0.9
VII	2.1

Table 8.4.1 ψ values for the seven transmembrane helices. Area of first peak is ignored in the calculation (see methods). The low values for helices II, III and VI suggest that these helices are not amphipathic.

The helical wheel representation and vertical plots (figure 8.4.1.5 to figure 8.4.1.7) depicting residue conservation in the transmembrane region of cationic ligand binding GPCRs (which includes the dopamine family of GPCRs) clearly shows that helices II and III are not amphipathic. Helix III lacks a Pro residue and so must fall into the category of being a regular helix (Barlow and Thornton, 1988). Helix III is also less exposed to the surrounding lipid (Baldwin, 1993) than for example helix IV and so arguably has more surface area either exposed to the protein interior or adjacent helices (II and IV). This suggests helix III has less scope for residue variation.

The consensus of opinion is that residues facing the surrounding lipid are free to undergo substitution mutations resulting in considerable variability. However the

variability profile of helix III particularly indicates low residue variability throughout the length of this transmembrane secondary structure. The situation with helix VI is less clear. There is a middle Pro residue in this helix and so falls into the category: non-regular helix (Barlow and Thornton 1988). However, helix V and helix VII both have middle Pro residues and both are clearly amphipathic in nature. Helix II has a conserved Pro residue at position 23 and so does not occupy a middle helix position. Consequently, the Pro residue of helix II is less likely to influence its ψ score. The possible role of helix II in helix interface shear mechanism with helix III is considered in the section 8.4.3.

8.4.2 Possible Role of Non-Amphipathic Helices in Binding Agonists

We would like to suggest that the lack of amphipathic character in helices II and III is of functional significance in terms of the binding of the natural agonist ligand. Saunders and Findlay (1990) has proposed a model (see figure 3.4.1; page 36) where the agonist binds first to the Asp residue on helix III and then binds to a deeper Asp residue on helix II (this is considered in more detail below). Maloney-Huss and Lybrand (1992) have clearly stated that helices are likely to undergo movements commensurate with the need to efficiently bind agonists.

8.4.3 Possible Role of Helix Interface Shear Mechanism in Binding Agonists

The concept of secondary structure motions was examined in a molecular dynamics simulation of the protein myohemerythrin[2] (Rojewska and Elber, 1990) who examined trajectories of helices. The fluctuations of the protein were found to be dominated by a rigid helix motions (RHM). The relative motions of these helices were found to be irregular, with no clear periodicity. Unfortunately the study ignored rotations about the long axes of the helices. However, in an earlier study Lesk and Chothia (1984) examined domain closure by comparing homologous helices in open and closed forms in citrate synthase (also for insulin: Chothia *et al.*, 1983). These workers

[2] Myohemerythrin (Klotz *et al.*, 1976) is an oxygen transport helix bundle protein composed of four helices, A-D.

showed that small shifts and helix rotations were accommodated not by changes in packing but rather by small conformational changes in side-chains which they called: *helix interface shear mechanism.* Shifts and rotations occurred at helix interfaces; rotations were in the range: 2^0 to 13^0 and shifts: 0.2 to 1.8 Å and that these movements are cumulative[3].

We would like to suggest that there is a strong likelihood that helix interface shear mechanism also occurs in the hepta-helical motif of the transmembrane region of GPCRs. In particular, the non-amphipathic transmembrane helices II and III are likely to undergo rotations. Possibly with transmembrane helix II rotating in an anti-clockwise fashion and transmembrane helix III in a clockwise fashion as viewed from the intracellular side of the membrane (for possible conformational changes in the TM region in response to agonist binding.see figures 8.4.3.1 to 8.4.3.3). Helix interface shear mechanism may allow conserved residues facing away from the protein interior of transmembrane helices II and III to come *into play* at the right moment during the agonist ligand binding process. Counter rotations of helices II and III would allow the agonist to equilibrate between the Asp residue on helix III and the deeper Asp residue of helix II. Alternatively, helix III might rotate clockwise in synchrony with helix II. Such rotations would tend to move Asp of helix III out of range of the cationic amine of the agonist while bringing Asp of helix II in range. In this way there would not be an equilibrium state between Asp on helix II and Asp on helix III in the exact manner hypothesised by Saunders and Findlay (1990).

It seems quite likely that evolution has sought to use helix shear in conjunction with helix kinking and other conformational changes caused by middle Pro residues in helix V, VI and VII. It is also possible that fluctuations in main-chain transmembrane helical hydrogen bonding pattern between 1.4 to 1.3, perhaps along just parts of the helix, may also provide an integral role in the agonist binding mechanism. For example, Chothia and Gerstein (1991) have noted that novel dynamics involving helix splitting into α-helical and 3_{10}-helical components plays an important role in the binding of lactate and

[3] Lesk and Chothia (1984) found that cumulative rotations can reach ~ 30^0.

NAD to lactate dehydrogenase. Answers to these questions will have to wait until high resolution structures exist for both bound and non-bound catecholamine binding GPCRs are available.

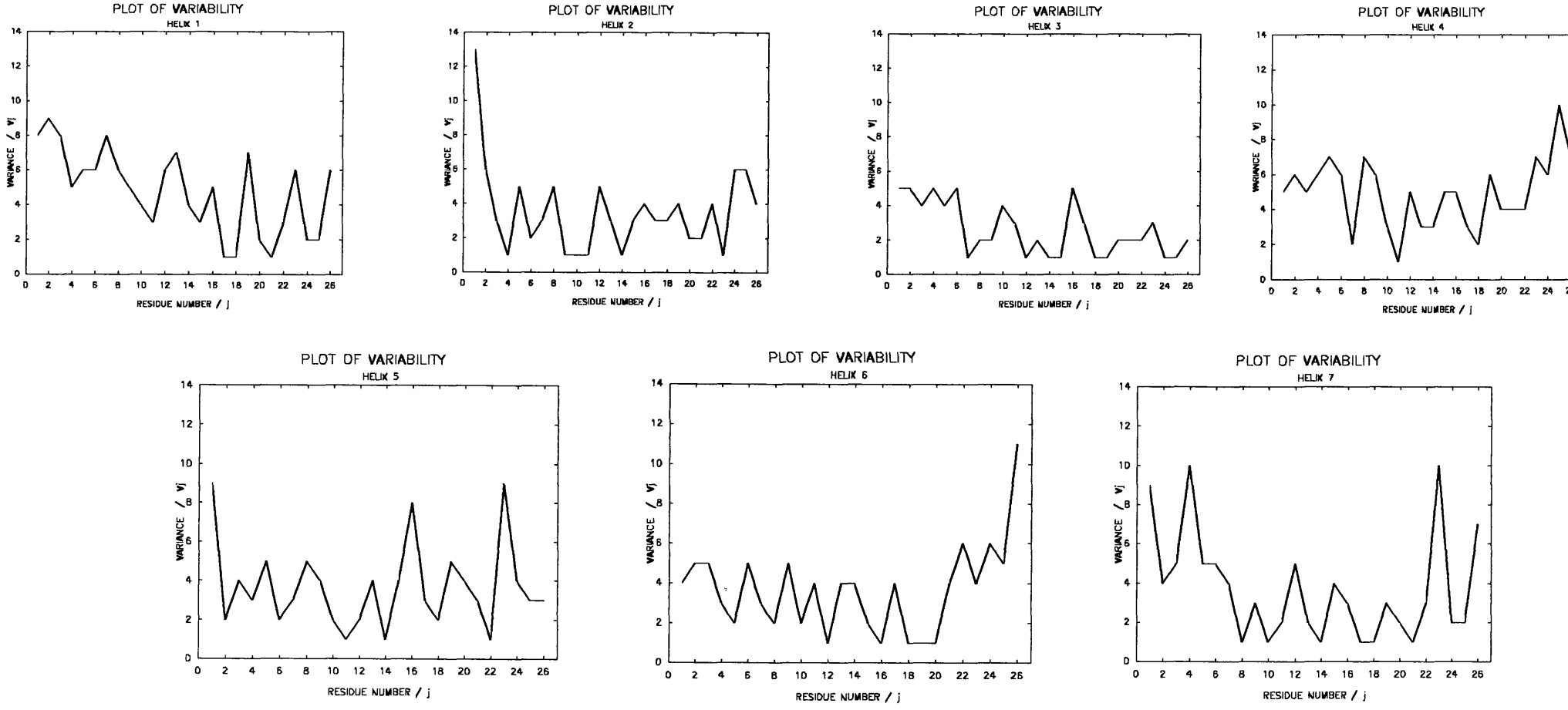


Figure 8.4.1.1 Variability plots (Donnelly *et al.*, 1989) produced by Peppi! for each putative transmembrane helix using a multiple sequence file (MSF; appendix 1) obtained using PILEUP command within Genetics Computer Group (GCG; Devereux *et al.*, 1984) software package and helix end positions correspond to those suggested by Baldwin, 1993. It is clear that putative transmembrane helices II, III and VI exhibit high conservancy (i.e. low variability) throughout their length - in particular helix III. Hence these helices are unlikely to amphipathic, i.e. unlikely that residues are conserved on one face of the helix only.

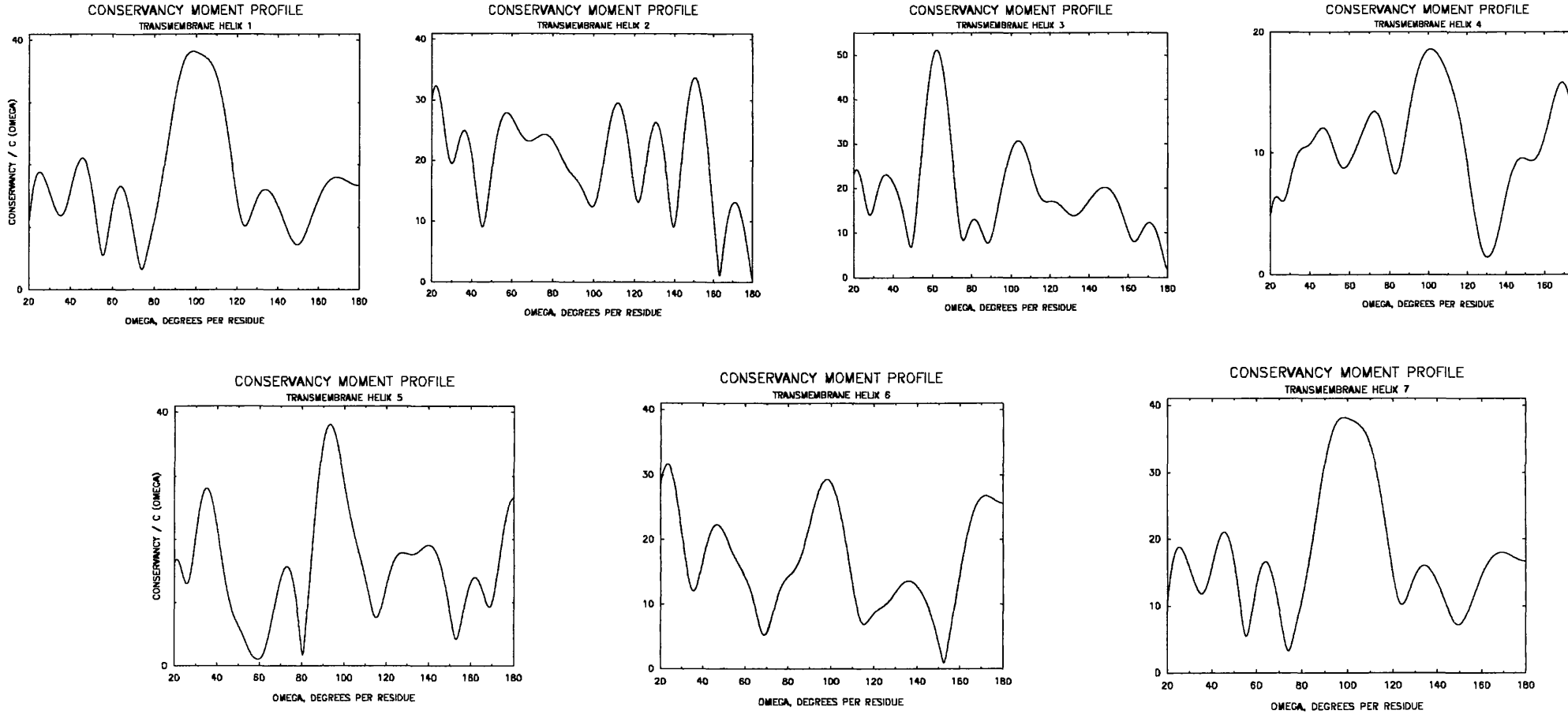


Figure 8.4.1.2 Conservancy moment profile plots (Eisenberg *et al.*, 1984) produced by Peppi! of each putative transmembrane helix using a multiple sequence file (MSF; appendix 1) obtained using PILEUP command within Genetics Computer Group (GCG; Devereux *et al.*, 1984) software package and helix end positions correspond to those suggested by Baldwin, 1993. Conservancy moment profile plots are able to detect amphipathic helices (peak at ~100^0) and amphipathic extended β-sheets (peak at ~180^0). Clearly helices I, IV, V and VII are amphipathic. Helices II and III are not amphipathic. Helix VI is borderline - it does have a peak at ~100^0 but it is not as dominant as the conservative moment profile with regard to helices I, IV, V and VII.

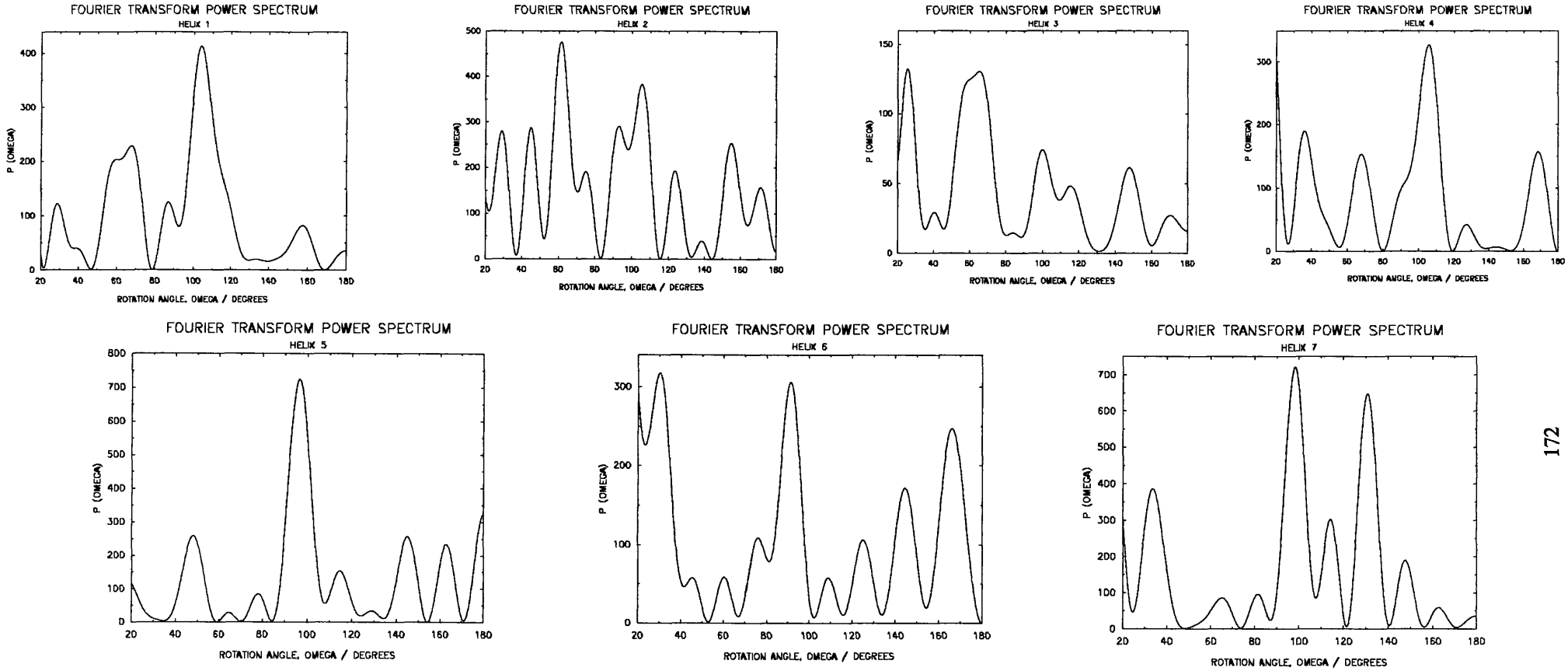


Figure 8.4.1.3 Fourier transform power spectra produced by Peppi! for each putative transmembrane helix using a multiple sequence file (MSF; appendix 1) obtained using PILEUP command within Genetics Computer Group (GCG; Devereux *et al.*, 1984) software package and helix end positions correspond to those suggested by Baldwin, 1993. It is clear that peaks occur at $\approx 100^0$ with regard to helices: I, IV, V, and VII. A peak at $\approx 100^0$ is expected for regular helices with conserved residues on one face only (i.e. 3.6 residues every $360^0 \equiv 100^0$). Hence these helices are amphipathic. Helices II and III are clearly not amphipathic. Putative transmembrane helix VI is difficult to interpret. This helix has a middle Pro residue and so is classified as an irregular helix according to the classification scheme suggested by Barlow and Thornton, 1988. However, helix V and VII both have a middle Pro residue and have dominant peaks at $\approx 100^0$. Helix V has a prominent peaks at $\sim 95^0$ (periodicity of d = 3.7 residues per turn) and corresponds to ω values obtained for amphipathic α-helices of water-soluble proteins (Cornette *et al.*, 1987); Yeates *et al.* (1987) has suggested that such peaks suggest a certain degree of left handed supercoiling resulting in slightly more residues per turn. Helix I and IV have prominent peaks at $\sim 105^0$ (d = 3.4). The increased value of ω reflects either a more tightly wound helix or a systematic shift in the residues that contact an adjacent helix, whose axis is tilted relative to the other helix axis (Komiya *et al.*, 1988). Helix VII has a $\omega \sim 100^0$ (d = 3.6) and is representative of a perfect α-helical conformation (albeit helix VII has a conserved middle Pro residue).

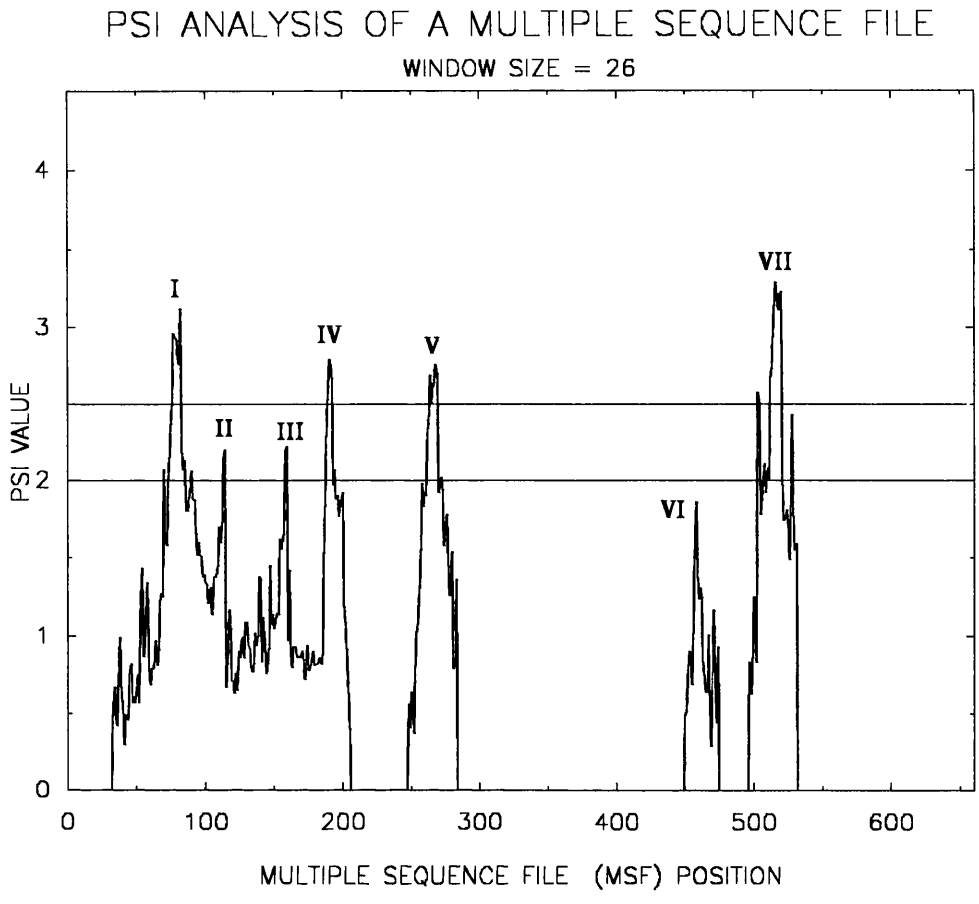


Figure 8.4.1.4 Psi analysis of a multiple sequence file (appendix 1) containing the dopamine, α and β-adrenergic families of G protein coupled receptors (GPCRs) together with the serotonin ($5HT_{1A}$) GPCR. Two horozontal lines are drawn. The first with a psi value of 2.0 which is the value recommended as a cut-off by Rees *et al.*, (1989b), suggests that there are seven amphipathic helices. However, given that each psi value was computed in a slightly different way (see methods section) requiring a less generous cut off value making 2.5 a better guide. Using a cut-off value of 2.5 clearly indicates that there are only four amphipathic helices: I, IV, V and VII. Helices II, III and VI are not amphipathic.

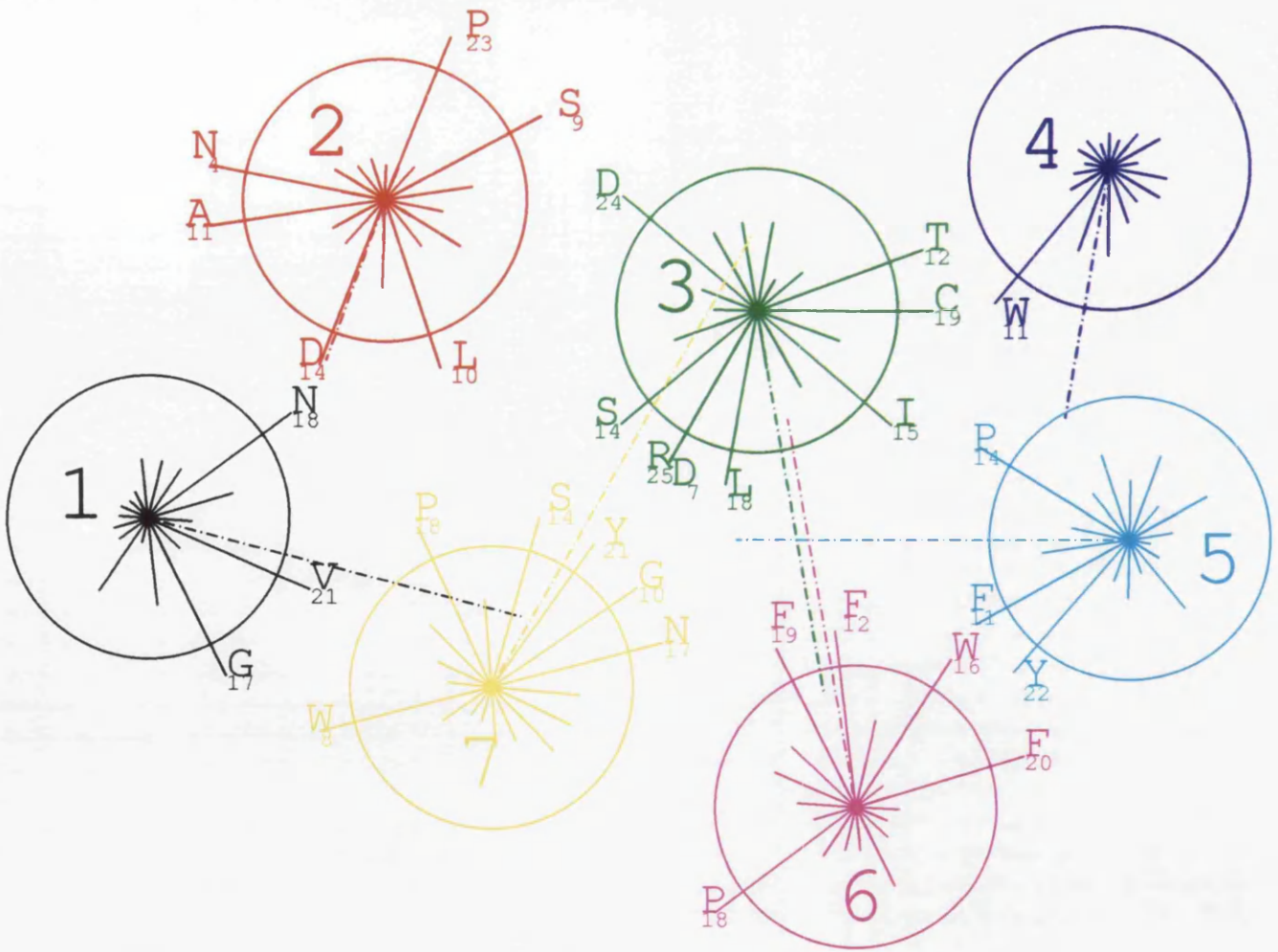


Figure 8.4.1.5 Helical wheel schematic produced by Peppi! Of the transmembrane region of the dopamine G protein coupled receptor as viewed from the intracellular side of the membrane. Dotted-dash lines represent the moment of conservancy for each helix; solid lines indicate relative conservancy. Only those residues which are conserved throughout the multiple sequence file (appendix 1) are labelled. Residue numbering (shown as subscript numbers) follows the recommendations of Baldwin (1993) - see figure 9.3.2 on page 189. It is clear that the helix II and III are not amphipathic. That is, residues are not conserved on one face of the helix, but are instead conserved on different faces of the same helix.

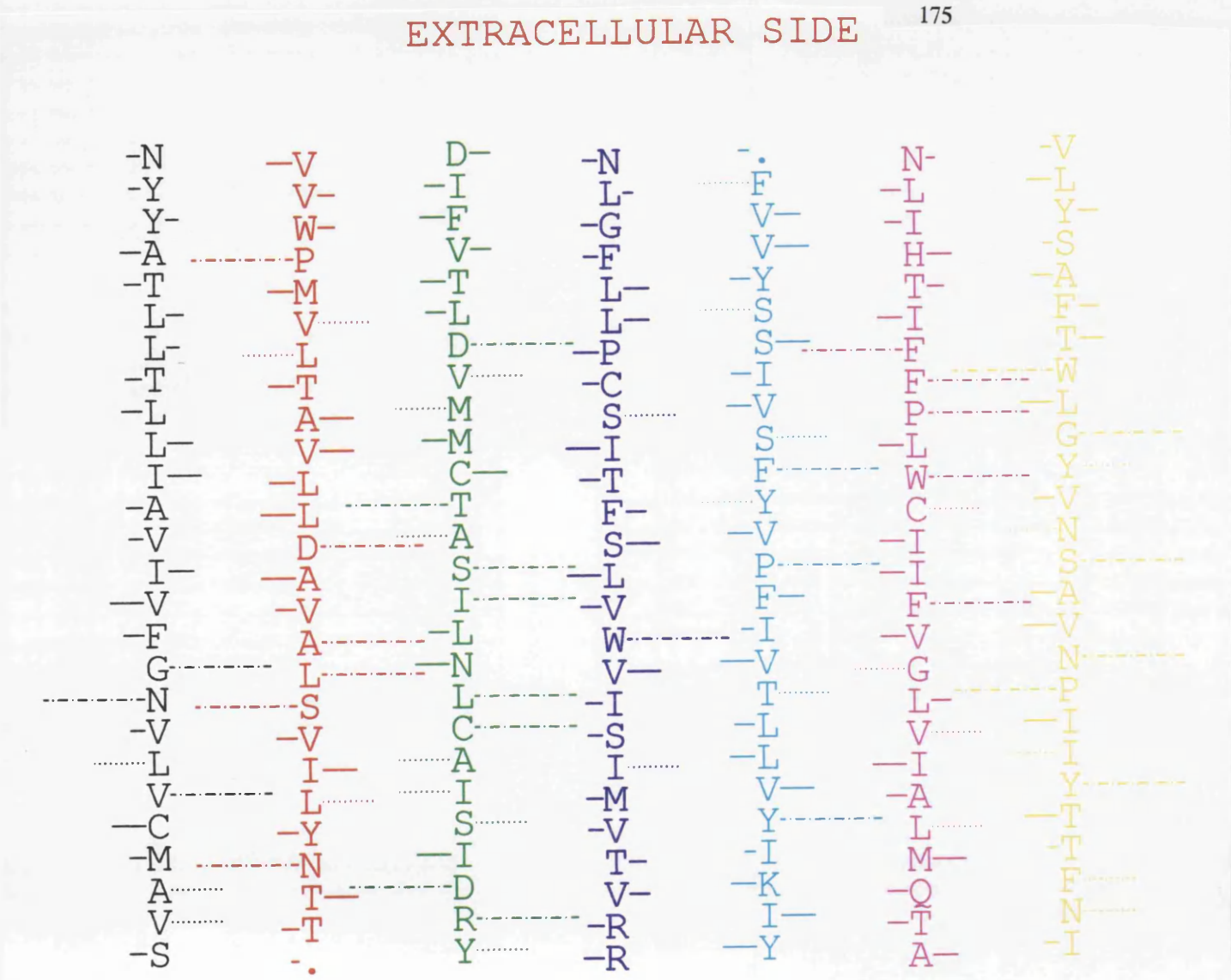


Figure 8.4.1.6 Vertical representation (Donnelly *et al.*, 1989) of residue conservation in each transmembrane helix (helix ends are the same as those used by Baldwin, 1993) based on multiple sequence file (appendix 1). Black = helix I; red = helix II, green = helix III and so on. Dotted-dashed lines indicate conserved residues. The right side of each vertical plot corresponds to the most conserved face (120^0) of each helix. Figure was produced using Peppi! coded by the author.

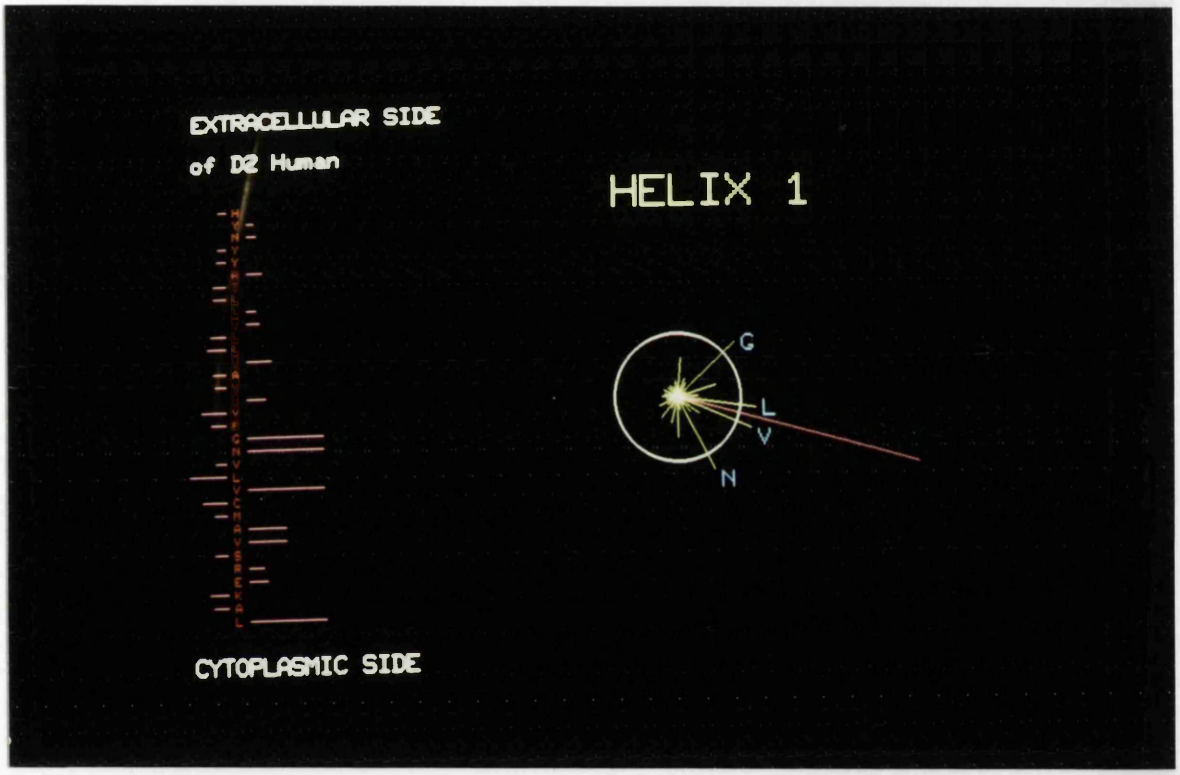
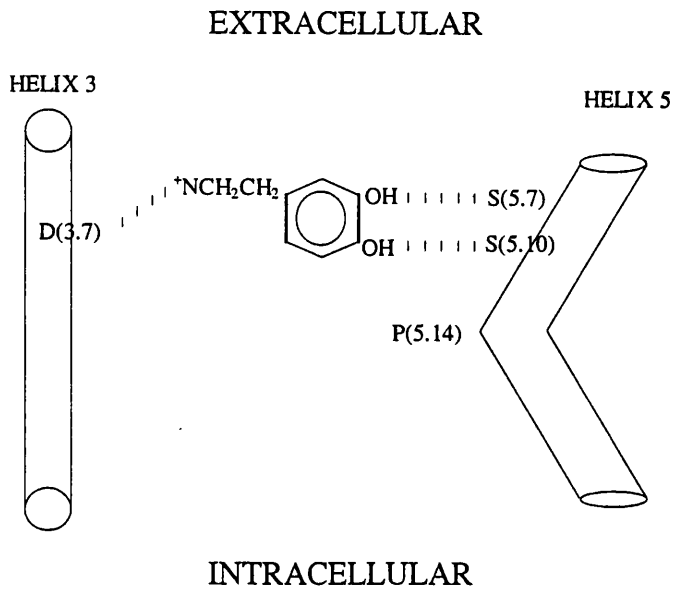


Figure 8.4.1.7 Picture of screen output combining both vertical and helical wheel representations of helix I. Helical wheel is depicted - this time looking end on from the extracellular side of the membrane. The vertical plot gives some indication of the variability along the helix.

(a)



(b)

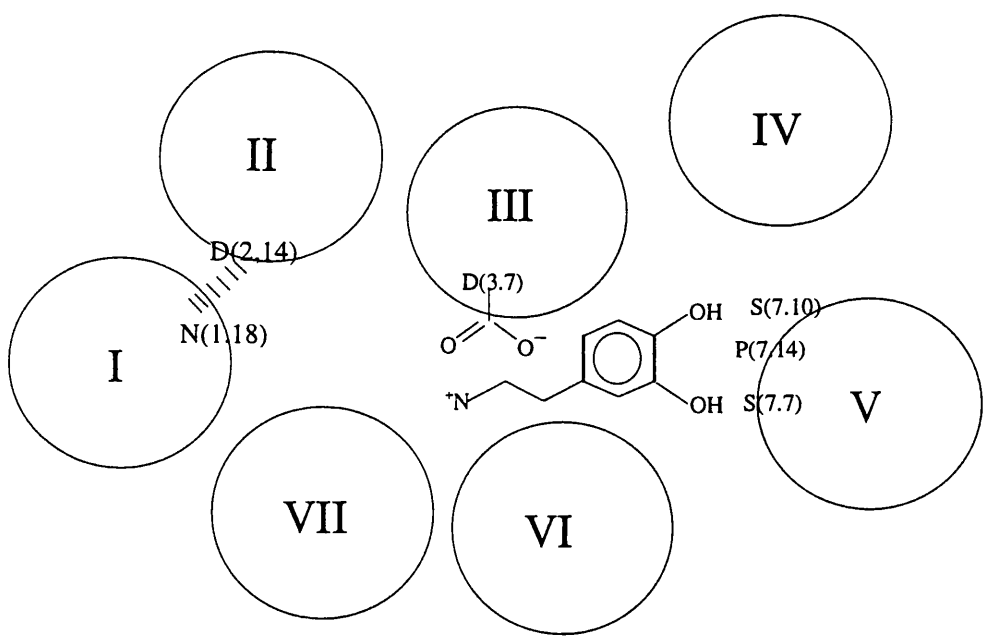
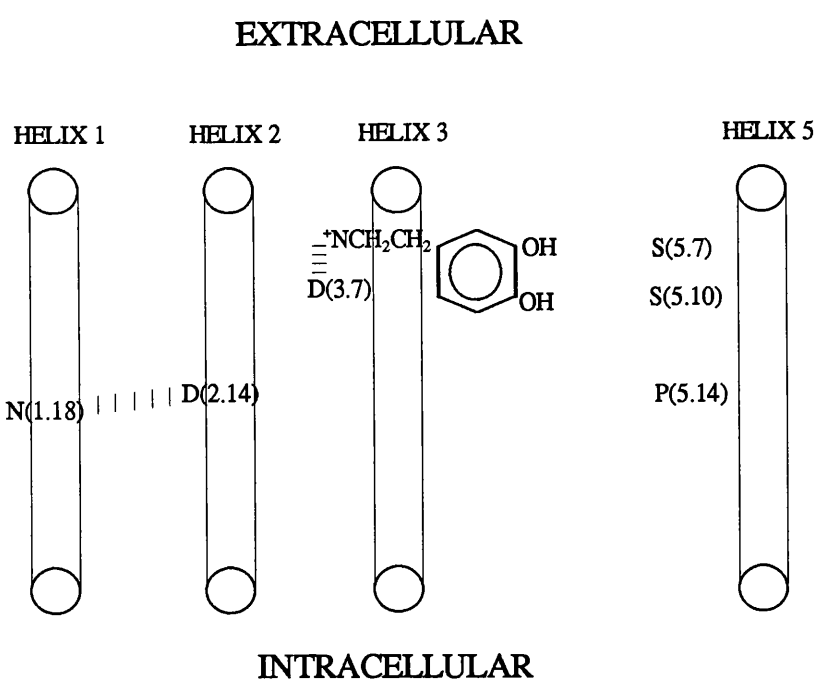


Figure 8.4.3.1 Possible conformational changes in the TM region in response to agonist binding. (a) Side view: Agonist is bound to D2 subtype dopamine receptor. The hydroxyl groups of the catechol ring of the natural ligand is shown hydrogen bonded to S(5.7) and S(5.10) and the cationic amine simultaneously forms hydrogen bonds and a reinforced ionic bond with D(3.7) - also see figure 3.4.1. The key feature exploited in this reinforced ionic bond is that it lasts longer than the weaker ionic bond which lasts approximately 10^{-5}s (Albert, 1971, 1979). It is also quite strong (up to 10 kcal/mol; Albert, 1971, 1979) with a separation distance of ≈ 3.5A [Vlijmen and IJzermann, 1989]. P(5.14) which is located in the middle of helix V is shown. Helix V is shown in its convex form with P(14.5) facing inwards towards the binding the interior of the binding pocket in the manner described by Gunnar (1991a). This initial phase in the binding of the agonist/antagonist has been referred to as Mode Ø (Saunders and Findlay, 1990); (b) Top view (from intracellular side of membrane): The interhelical bond between N(1.18) and D(2.14) is shown though it does not play a direct part in Mode Ø. Legend otherwise the same as for (a).

(a)



(b)

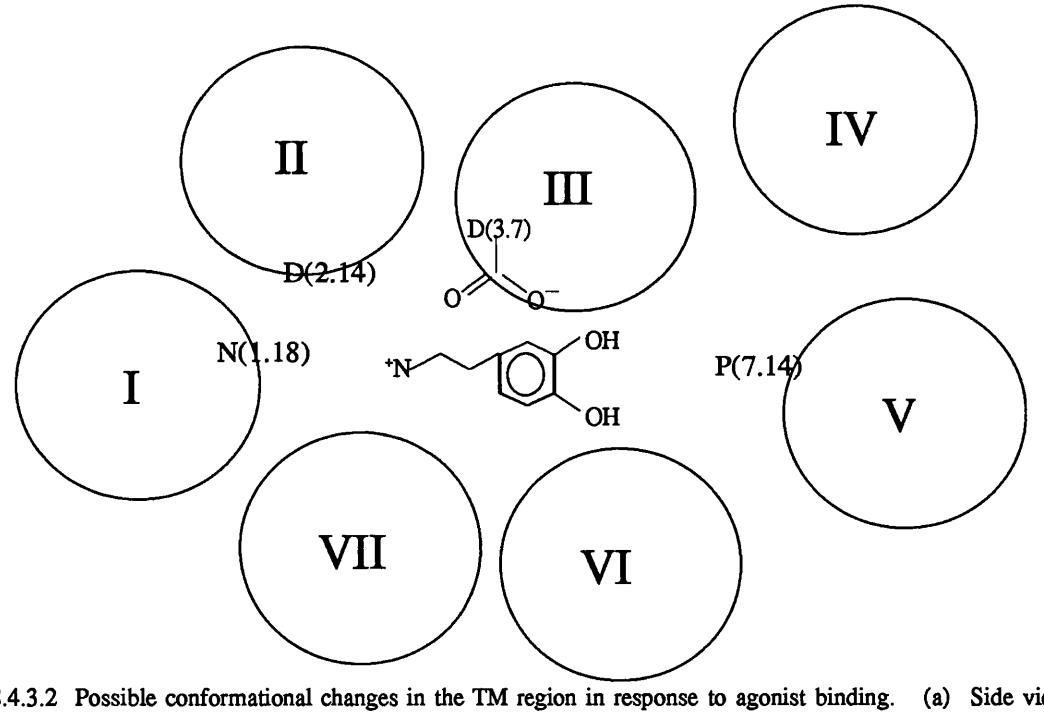


Figure 8.4.3.2 Possible conformational changes in the TM region in response to agonist binding. (a) Side view: P(5.14) in the middle position of helix 5 causes kinking on a pico-second time scale [Sankararamakrishnan *et al.*, 1991]; the kink angle can be as much 50^0 (MD average: 28.5^0 ($\pm 11.7^0$)) and wobble angle in the range -50^0 to 90^0 (MD average: 22.9^0 ($\pm 28.8^0$)) [Sankararamakrishnan *et al.*, 1991]. These flexing motions initially facilitate the agonist in its motion past helix 3 while maintaining a reinforced ionic bond with D(3.7). This signals the first agonist-specific binding action (the molecular dynamics output of helix 5 is wasted on antagonists). To accommodate the movement of the cationic amine, helix 3 rotates in sympathy about its axis in a clock-wise cork-screw by means of the helix shear mechanism (see main text). The helix shear is transmitted from helix 3 directly to helix 2 which rotates in the opposite direction (counter-clockwise). This helix shear in turn weakens the hydrogen bond between N(1.18) located in helix I and D(2.14) located in the lower half of helix 2 (relative to the extracellular region). The continuing motion of the agonist is towards D(2.14) and is facilitated by favourable σ-π interactions between the agonist amine and surrounding aromatic residues. In this proposal it does not seem possible to classify any part of this binding process of the agonist as Mode 1. (b) top view (from intracellular side of membrane); legend otherwise same as for (a).

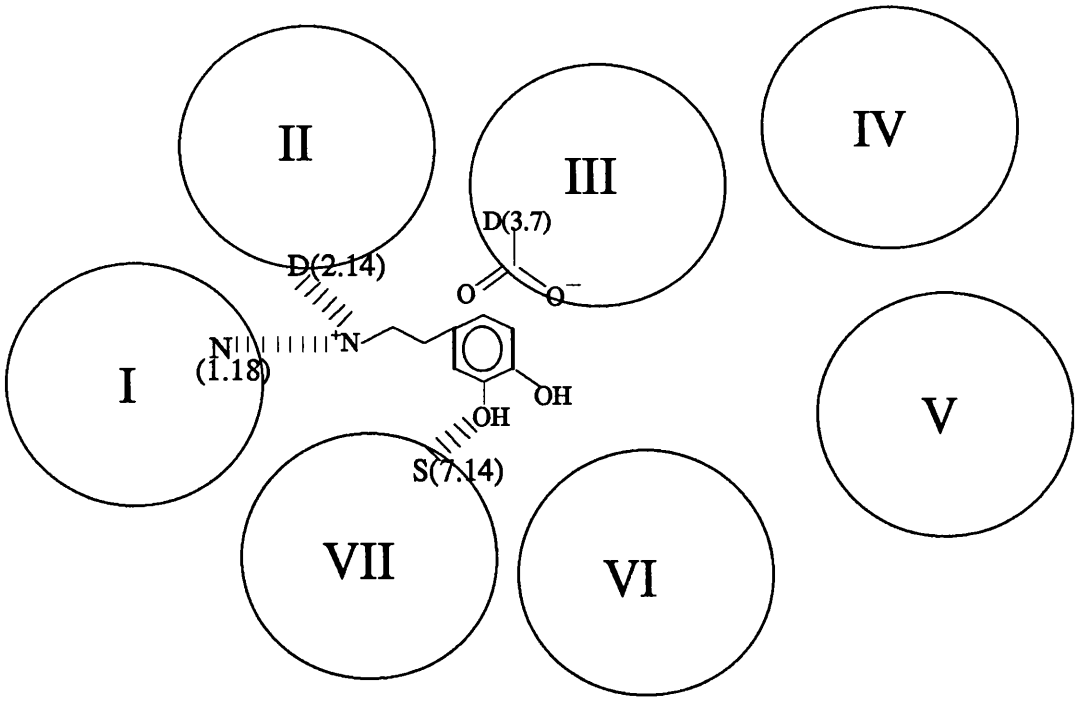


Figure 8.4.3.3 Top view as seen from intracellular side of membrane: The agonist is now in a position to form a reinforced ionic bond with D(2.14) - this is aided by the flexing action of helix VII due to the conserved P(7.18). The side chain of N(1.18) may be involved in a hydrogen bonding to the amine of the agonist or alternatively may point directly at the center of the aromatic ring of the agonist in the same manner as reported by Perutz *et al.* (1986) and reviewed by Levitt and Perutz (1988). S(7.14) is highly conserved in the catecholamine binding class of GPCRs and is therefore shown hydrogen bonded to a catechol hydroxyl group. The cork-screw action of helix III has exposed the DRY motif at the carboxy-terminal end to the intracellular domain. Helix III is held in this state while the cationic amine is held at a separation distance of ≈ 3.5A from D(2.14). G protein coupling involving iII and iIII is now able to take place.

8.5 Conclusion

Detailed Fourier analysis clearly suggest that the consensus view that only those residues facing the transmembrane protein interior of GPCRs are conserved is flawed. This study has shown that helices II and III are not amphipathic. We suggest that this lack of amphipathicity is extremely important in aiding agonist ligands to bind to a second Asp residue on helix II of catechol amine binding GPCRs - in particular the binding of agonists to D_1 and D_2 subtype dopamine receptors.

9. A Molecular Modelling Study Of The Dopamine Family Of G Protein Coupled Receptors (D_1, D_2, D_3, D_4 And D_5).

9.1 SUMMARY

The intermediate resolution map of bacteriorhodopsin is frequently used as a template to model catechol amine binding G protein coupled receptors. However, it has been proposed that such modelling studies are likely to be misdirected due to lack of sequence homology in the seven helix bundle region between bacteriorhodopsin and any G protein coupled receptor. We have noted that such models frequently pay insufficient attention to the modelling of kinks caused by Pro residues, particularly where they occupy middle positions in transmembrane helices. In particular, the probable influence of the middle Pro in transmembrane helix V (TM5) on local molecular dynamics is largely ignored at the expense of a fuller understanding of the likely mechanism of agonist binding. Likewise, interhelical hydrogen bonding is rarely discussed in the few modelling studies which have been published to date. The molecular details of the three dimensional models are very difficult to compare and interpret due to differences in the procedures used in model building and the unavailability of final atomic co-ordinates. Previously published modelling studies of G protein coupled receptors have made extensive use of energy minimisation. Several workers have noted that extensive use of energy minimisation causes compaction. Here we present a modelling approach which has sought as far as possible with the resources available to be sensitive to each of these issues.

9.2 INTRODUCTION

On the basis of homology and pharmaceutical studies the dopamine family of catechol amine binding G protein coupled receptors (GPCRs) clearly fall into two subfamilies: D_1 and D_2. The D_1 subgroup currently includes D_1 and D_5 (Dearry, *et al.*, 1990 and Sunahara *et al.*, 1991 respectively); the D_2 subgroup: D_2, D_3 and D_4 (Dal-Toso *et al.*, 1989; Sokoloff *et al.*, 1990 and Van Tol *et al.*, 1991 respectively). The DA

receptors are targets of drug therapy in disorders such as Parkinson's disease and schizophrenia.

With few exceptions 3D modelling of GPCRs have used bacteriorhodopsin (bRh) as a structural template even though there is no homology (chapter 5). Notable exceptions being the 3D modelling of the entire β_2 receptor (Maloney-Huss and Lybrand, 1992) and the serotonin 5-HT_{1A} receptor (Sylte *et al.*, 1993) - both of which were modelled *de-novo* and so explicitly avoided using bRh as a template. More recently, knowledge-based modelling approaches have been developed such as the modelling of transmembrane (TM) seven helix bundles (Cronet *et al.*, 1993). However, each of these approaches use different helix end positions and helix phase. Also, while each of these approaches incorporates the same overall topography (i.e. hepta-helical TM motif with helices arranged in an anti-parallel fashion[1]) the actual spatial arrangement of the helices in each model is significantly different. Hence, this adds credence to the notion that 3D models of GPCRs are difficult to compare - a point strongly argued by Humbler & Mizadegan, 1992.

A very detailed prediction of helix phase, orientation and helix ends is already available (Baldwin, 1993) and makes use of the latest projection map of rhodopsin (rH) - Gebhard, *et al.*, 1993. The probable arrangement of the helices in GPCRs as discerned by Baldwin is easily applied to any cloned member of this supergene class of receptor proteins. It makes good sense for the wider GPCR modelling community to use this respected study to generate 3D models. In this way, models produced by different groups will be easier to compare and interpret. Therefore, the start and end points, helix phase, approximate tilts and positions of individually modelled TM helices used in the final models are all based on Baldwin's conclusions:

- membrane spanning segments are α-helices arranged sequentially in an anti-parallel fashion
- each TM helix contains 26 amino acids
- the most conserved face of each TM helix faces towards the interior of the receptor protein

[1] Transmembrane helices I and VII are parallel to one another.

- helices I, IV and V (TM1, TM4 and TM5) are more exposed to the surrounding lipid
- the arrangement of the helices is similar to that revealed by the latest 2D experimental projection map obtained for rH

It is well known that Pro residues located in TM helices are highly conserved, cause kinks and are important in the functioning of the receptor. TM5 contains a middle Pro which is conserved in every member of the catecholamine and muscarinic acetylcholine binding families of GPCRs. While modelling studies frequently refer to a helix-Pro residue functioning as a *hinge,* two recent studies have characterised the molecular dynamics of a TM helix with a middle *trans*-Pro with particular emphasis on pyrrolidine ring puckering and its relationship with backbone dihedral torsion angles of preceding residues (Sankararamakrishnan and Vishveshwara, 1990; Sankararamakrishnan *et al.,* 1991). It is evident from these molecular dynamics studies that TM5 is likely to be oscillating between a largely straight structure and a bent structure every 2-4ps. Also that TM5 will tend to orient itself with its convex sides towards the receptor interior and concave side towards the surrounding lipid in line with the conclusions of Gunnar (1991a) who has studied the impact of Pro kinks in TM α-helices. The likely role of the middle Pro in TM5 in guiding agonist ligands deeper into the binding pocket has not been adequately discussed in the literature to date though Williams and Deber (1991) argue that the function of Pro in TM helices is to provide rigidity. However, their analysis was influenced by structural and modelling studies of bRh where the ligand (a retinal chromophore) is permanently, though reversibly, bound.

Since the cyclic side chain of Pro places constraints on the backbone dihedral angles both of itself and the preceding residue (Carver and Blout, 1967; Deber *et al.,* 1990) the backbone ϕ, ψ angles of XProY were examined in a study by Polinsky *et al.* (1992)[2] who derived minimum energy conformations of Pro containing helices in a membrane environment. The *Trans*-I conformation being the most populated lowest energy family, corresponding to a Pro in a kinked α-helix in a membrane environment. The average *Trans*-I ϕ, ψ dihedral

[2] Williams and Deber are co-authors of this paper.

torsion angles for the three residues XProY being: {-50.6°,-53.1°}, {-75.0°, -39.4°} and {-72.9°,-38.3°} respectively[3]. Though the energy barrier separating these conformations was not investigated it seems likely that a TM helix with a middle Pro will sample various low energy conformational states on a pico-second time scale in line with the slightly earlier findings of Sandararamakrishnan *et al.*, 1991.

Reid and Thornton (1989) showed that about 40% of the χ_1 torsional angles can be modelled correctly if statistically preferred values are used. Therefore, the model side-chain torsional angles can be adjusted to reflect the primary values observed in a statistical studies of side chain torsional angles as a function of 2° structure carried out by McGregor *et al.* (1987) and Sutcliffe *et al.*, (1987). McGregor *et al.* (1987) looked at χ_1 and χ_2 torsional angles as a function of secondary structure and position along the α-helix: α-helix (centre), α-helix (N end) and α-helix (C end). Sutcliffe *et al.* (1987) extended the work of McGregor *et al.* (1987) to suggest "best shot" χ_3 and χ_4 torsional angles. The rotamer library developed by Ponder and Richards (1987), which is frequently quoted in the literature, is of limited value since it failed to take account of secondary structure - χ values were statistically derived using whole tertiary structures. Since the TM region of GPCRs is primarily a seven TM helix motif, it is obviously important to use the preferred χ torsion angles for α-helices in modelling work.

While hydrogen bonding between ligands and TM regions is frequently discussed it is noticeable that interhelical hydrogen bonding rarely receives attention. The reason for this is not clear even though TM helices frequently contain interhelical hydrogen bonds - though the average is less than one interhelical hydrogen bond per TM helix (Lemmon and Engelman, 1991). For example, in the intermediate resolved structure of bRh (Henderson *et al.*, 1990), it appears that Asp212 in helix G (TM7) is involved in hydrogen bonding to Tyr57 of helix B(TM2), Trp86 of helix C (TM3) and to Tyr185 of helix F (TM6). Also, charged side-chains located in TM helices are generally assumed to be involved in non-bonded interactions with one or more counter ions (Engelman *et al.*, 1980). One notably exception being in the case of a recent modelling study of β_2 where it was noted that the glutamic acid side chain (GLU122) is exposed to the lipid, although it was also thought that

[3] The ϕ torsion angle of Pro was not quoted explicitly by Polinsky *et al.* (1992).

it could form a hydrogen bond with the exposed threonine (THR118) - Maloney-Huss and Lybrand, 1992. A recent study of the TM regions of photosynthetic reaction centre proteins showed that highly conserved tryptophan residues are likely to hydrogen bond to a carbonyl oxygens in other TM helices (Schiffer, *et al.*, 1992).

9.3 METHOD

Essentially, the combinatorial approach (Cohen *et al.*, 1979) to protein modelling was used - figure 9.3.1. The TM regions of each member of the dopamine family of GPCRs (human D_1, D_2, D_4 and D_5 dopamine receptors and rat D_3 dopamine receptor) was modelled using start and end residue positions for each TM helix in accordance with the recommendations of Baldwin (1993) - figure 9.3.2. Phi/psi angles used were -59°/-44° for helical residues in a non-polar environment (Blundell *et al.*, 1983). With the exception of TM5, kinks in TM helices caused by middle Pros were modelled using the phi/psi angles based on the *Trans*-I conformation, corresponding to a Pro in a kinked α-helix in a medium of low polarity (Polinsky *et al.*, 1992). TM5 was modelled using backbone starting co-ordinates, corresponding to a helix with a Pro occupying the middle position and its pyrrolidine ring in a *puckered down* conformation (table 9.3.1). Side-chain torsion angles were modelled using statistically derived values based on secondary structure (Sutcliffe *et al.*, 1987); table 9.3.2.

The energy of each modelled helix was calculated using the colour force option of SYBYL[4] to allow easy identification of bad side-chain contacts. Side-chain clashes were removed by adjusting χ_1 torsion angles in line with probable values based on secondary structure (McGregor *et al.*, 1987). Where bad contacts remained the side-chain torsion SCAN within SYBYL was used to achieve satisfactory side-chain geometry. Individual TM helices were then energy minimised to convergence (0.1 kcal/mol) by conjugate gradient method using the Kollman all-atom force field (Weiner *et al.*, 1986) and Kollman charges - see appendix 4. The dielectric constant was set to 5.0 to take account of the non-

[4] All modelling and energy minimisations were carried out on an Evans & Sutherland 10/33 workstation running SYBYL (version 6.0); both products supplied by Tripos Inc., St. Louis, U.S.A. The energy minimisation module built into SYBYL is a user friendly mini-expert system allowing the user to set up molecular mechanics simulations very quickly.

polar environment. Non-bonded interactions cut-off and the 1-4 scaling factor were set to 10Å and 0.5 respectively.

The seven TM regions of D_2 were assembled to mimic the probable arrangement of helices in GPCRs as described by Baldwin (1993). Firstly, TM2 was manually docked to TM1 taking care to maintain appropriate helix phase (figure 9.3.3) and relative tilts (figure 9.3.4). Then the total energy of this pair of helices was calculated using the Kollman united-atom force field (Weiner *et al.*, 1984), which was used for all subsequent energy minimizations. Slight adjustment of the helix positions was allowed to remove excessive overlap. Bad side-chain contacts between the helices were manually removed by adjusting χ_1 torsion angle of side-chains again using statistically derived values based on secondary structure (McGregor *et al.*, 1987). Where necessary, χ_2 or χ_3 torsion angles were also adjusted to alternative values (±60°, 180°). Side-chain SCAN option of SYBYL was not used since this option does not take into account bad-contacts between different secondary structures. When all bad contacts likely to dominate subsequent energy minimisation were removed, the helix pair was subjected to just 20 iterations of energy minimisation

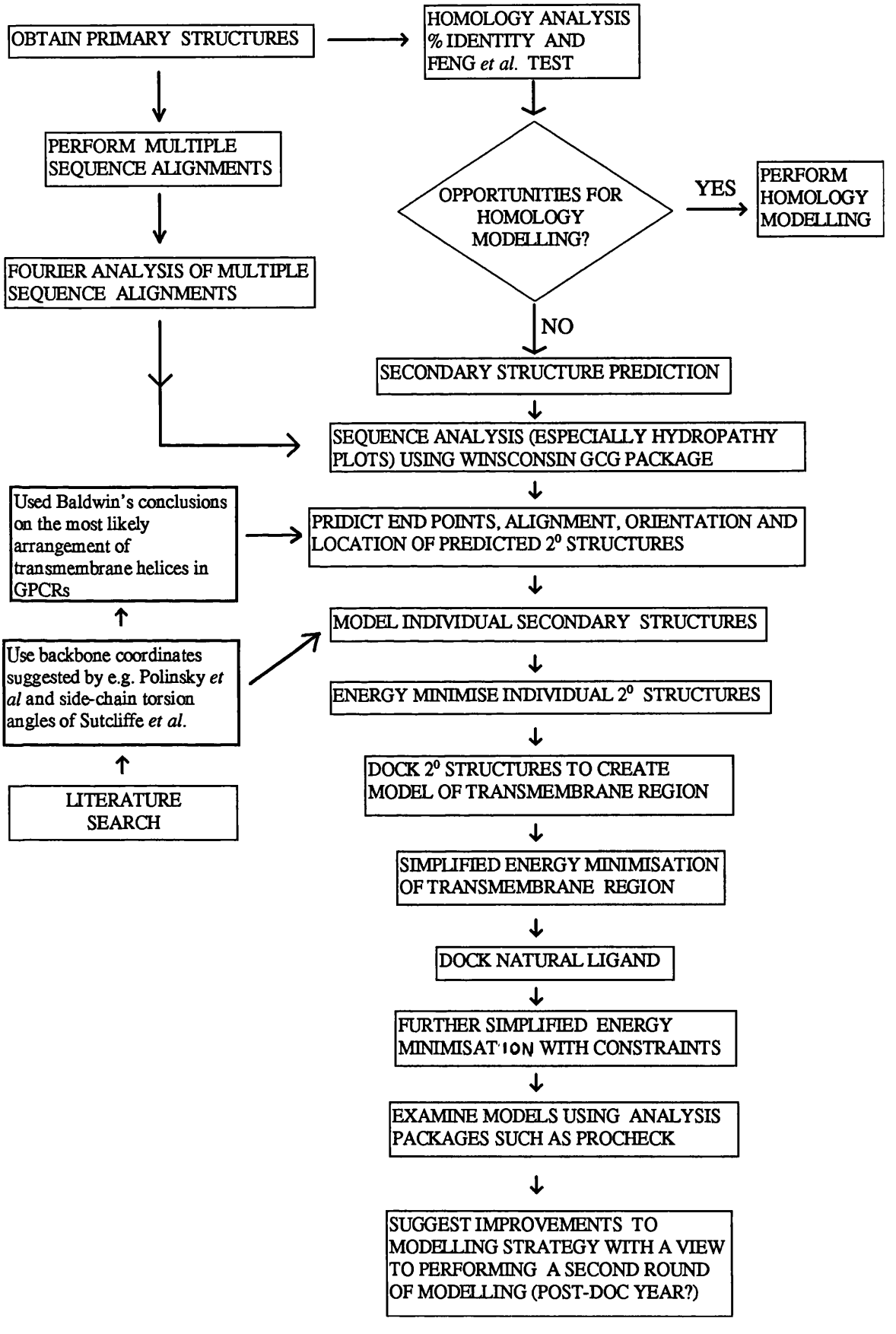
Likewise, TM3 was docked to TM2 (keeping TM1 and TM2 stationary), bad contacts were removed and all three helices were subject to 20 iterations of energy minimisation. TM4 was then docked to TM3 and so on. Finally, TM7 was docked to both TM1 and TM7 and the TM region was subjected to just 50 iterations of energy minimisation to establish that the structure was indeed stable and not subjected to compaction. Procheck (Laskowski, *et al.*, 1993) was used to check the steriochemical quality of the models (ϕ/ψ and χ_1/χ_2 plots). Surfaces and volumes were analysed using GRASP (version 1.1; Nicholls *et al.*, 1993). The whole procedure was repeated to model the TM region of human muscarinic M_1 (Peralta *et al.*, 1987) receptor.

The TM model of D_2 was then used as a template to model the remaining members of the D_2 sub-family (i.e. D_3 and D_4 subtypes). Similarly a separate model was constructed for D_1 which was used as a template to model D_5. This did not involve multiple site-directed mutagenesis. Instead, each TM helix was modelled separately and energy minimised as described above. Then, for example, the TM region of D_3 was constructed by using

backbone atoms to superimpose each energy minimised TM helix onto corresponding TM helices of D_2 (prior to the docking of the agonist ligand). The whole TM region was then examined for bad contacts using the colour force option of SYBYL as explained above. The TM region was then subjected to just 100 iterations of energy minimisation using the Kollman united atom force field and charges. The object being to create a physically plausible structure albeit without interconnecting helix loops, amino and carboxy terminal sequences.

The molecular structure of dopamine (Giesecke, 1980) was obtained using the Cambridge Structural Database (Allen and Kennard, 1983), energy minimised using the Tripos proprietary force field (charges calculated using the Geister-Huckel method as implemented within SYBYL) and manually docked to the binding pocket of D_2. Slight distance constraints were then applied between *m*-hydroxyl group of the catechol ring and Ser_{507} and likewise for *p*-hydroxyl group and Ser_{510} and the whole complex was subjected to 50 iterations in lots of 10 using the Tripos force field and Gasteiger-Huckel calculated charges. [The first digit of the subscript number is the TM number and the next 2 digits represents the position of the residue in the TM helix. Hence, Ser_{507} is the 7th residue in TM5. A similar numbering scheme was used by Trumpp-Kallmeyer *et al.*, (1992).]

Figure 9.3.1 - overleaf. Flow chart summarising the various stages used to develop 3D models of the transmembrane region of D_2. The first stages involved assessing the possible use of bRh as a structural template for homology modelling (chapter 5). The combinatorial modelling approach (Cohen *et al.*, 1979) was subsequently used for the first round of modelling, i.e. secondary structure prediction and characterisation (chapters 5 and 8). However, this approach then developed into a knowledge-based modelling approach in the sense that the literature supplied structural templates in the form of the most probable arrangement of helices in GPCRs (Baldwin, 1993), the most likely conformation of kinked helices (Sankararamakrishnan and Vishveshwara, 1990; Sankararamakrishnan *et al.*, 1991; Polinsky *et al.* 1992) and the most probable side-chain torsion angles (Sutcliffe *et al.*, 1987). Overall, the modelling approach took the form of an ad-hoc hybrid approach or "by hook or by crook" approach as described by Wishard and Muir (1990). However, if a new set of GPCR structures was required the modelling would immediately start with the work of Baldwin.


OBTAIN PRIMARY STRUCTURES
HOMOLOGY ANALYSIS % IDENTITY AND FENG et al. TEST
PERFORM MULTIPLE SEQUENCE ALIGNMENTS
OPPORTUNITIES FOR HOMOLOGY MODELLING?
YES
PERFORM HOMOLOGY MODELLING
FOURIER ANALYSIS OF MULTIPLE SEQUENCE ALIGNMENTS
NO
SECONDARY STRUCTURE PREDICTION
SEQUENCE ANALYSIS (ESPECIALLY HYDROPATHY PLOTS) USING WINSCONSIN GCG PACKAGE
Used Baldwin's conclusions on the most likely arrangement of transmembrane helices in GPCRs
PRIDICT END POINTS, ALIGNMENT, ORIENTATION AND LOCATION OF PREDICTED 2⁰ STRUCTURES
MODEL INDIVIDUAL SECONDARY STRUCTURES
Use backbone coordinates suggested by e.g. Polinsky et al and side-chain torsion angles of Sutcliffe et al.
ENERGY MINIMISE INDIVIDUAL 2⁰ STRUCTURES
LITERATURE SEARCH
DOCK 2⁰ STRUCTURES TO CREATE MODEL OF TRANSMEMBRANE REGION
SIMPLIFIED ENERGY MINIMISATION OF TRANSMEMBRANE REGION
DOCK NATURAL LIGAND
FURTHER SIMPLIFIED ENERGY MINIMISATION WITH CONSTRAINTS
EXAMINE MODELS USING ANALYSIS PACKAGES SUCH AS PROCHECK
SUGGEST IMPROVEMENTS TO MODELLING STRATEGY WITH A VIEW TO PERFORMING A SECOND ROUND OF MODELLING (POST-DOC YEAR?)

```
                    HELIX I      (TM1)          HELIX II (TM2)
D1 Human   ILTACFLSLLILSTLLGNTLVCAAVI KVTNFFVISLAVSDLLVAVLVMPWKA
D2 Human   NYYATLLTLLIAVIVFGNVLVCMAVS TTTNYLIVSLAVADLLVATLVMPWVV
D3 Rat     AYYALSYCALILAIIFGNGLVCAAVL TTTNYLVVSLAVADLLVATLVMPWVV
D4 Human   AAALVGGVLLIGAVLAGNSLVCVSVA TPTNSFIVSLAAADLLLALLVLPLFV
D5 Human   VVTACLLTLLIIWTLLGNVLVCAAIV NMTNVFIVSLAVSDLFVALLVMPWKA
           12345678901234567890123456 12345678901234567890123456

                    HELIX III (TM3)          HELIX IV (TM4)
D1 Human   NIWVAFDIMCSTASILNLCVISVDRY KAAFILISVAWTLSVLISFIPVQLSW
D2 Human   DIFVTLDVMMCTASILNLCAISIDRY RRVTVMISIVWVLSFTISCPLLFGLN
D3 Rat     DVFVTLDVMMCTASILNLCAISIDRY RRVALMITAVWVLAFAVSCPLLFGFN
D4 Human   DALMAMDVMLCTASIFNLCAISVDRF RRQLLLIGATWLLSAAVAAPVLCGLN
D5 Human   DVWVAFDIMCSTASILNLCVISVDRY RMALVMVGLAWTLSILISFIPVQLNW
           12345678901234567890123456 12345678901234567890123456

                    HELIX V (TM5)            HELIX VI (TM6)
D1 Human   TYAISSSVISFYIPVAIMIVTYTRIY VLKTLSVIMGVFVCCWLPFFILNCIL
D2 Human   AFVVYSSIVSFYVPFIVTLLVYIKIY ATQMLAIVLGVFIICWLPFFITHILN
D3 Rat     DFVIYSSVVSFYVPFGVTVLVYARIY ATQMVVIVLGAFIVCWLPFFLTHVLN
D4 Human   DYVVYSSVCSFFLPCPLMLLLYWATF AMRVLPVVVGAFLLCWTPFFVVHITQ
D5 Human   TYAISSSLISFYIPVAIMIVTYTRIY VLKTLSVIMGVFVCCWLPFFILNCMV
           12345678901234567890123456 12345678901234567890123456

                    HELIX VII (TM7)
D1 Human   NTFDVFVWFGWANSSLNPIIYAFNAD
D2 Human   VLYSAFTWLGYVNSAVNPIIYTTFNI
D3 Rat     ELYRATTWLGYVNSALNPVIYTTFNV
D4 Human   RLVSAVTWLGYVNSALNPVIYTVFNA
D5 Human   TTFDVFVWFGWANSSLNPVIYAFNAD
           1234567890123456789012345
```

Figure 9.3.2; Helix assignments of the dopamine family of G protein coupled receptors. Transmembrane (TM) assignments follow those of Baldwin (1993).

Dihedral angles	Average Value (degrees)
$\phi_{\rho -5}$	-47.05
$\psi_{\rho -5}$	-42.41
$\omega_{\rho -5}$	-179.80
$\phi_{\rho -4}$	-57.52
$\psi_{\rho -4}$	-29.75
$\omega_{\rho -4}$	175.20
$\phi_{\rho -3}$	-90.50
$\psi_{\rho -3}$	-48.68
$\omega_{\rho -3}$	-175.83
$\phi_{\rho -2}$	-53.71
$\psi_{\rho -2}$	-47.88
$\omega_{\rho -2}$	-178.70
$\phi_{\rho -1}$	-56.15
$\psi_{\rho -1}$	-63.85
$\omega_{\rho -1}$	178.41
ϕ_{ρ}	-59.99
ψ_{ρ}	-40.13
ω_{ρ}	177.43
$\phi_{\rho +1}$	-56.10
$\psi_{\rho +1}$	-49.94
$\omega_{\rho +1}$	178.73
$\phi_{\rho +2}$	-63.35
$\psi_{\rho +2}$	-44.58
$\omega_{\rho +2}$	178.12
χ_1	21.90
χ_2	-31.90

Table 9.3.1; Dihedral angles for α-helical structure with a middle Pro in the *puckered-down* conformation (i.e. χ_1 is positive and χ_2 is negative). Derived from 11 structures corresponding to 75-85ps period of a 100ps molecular dynamics simulation during which the Pro was in the *puckered-down* conformation. Average values obtained following Newton-Raphson minimization. These backbone parameters were used to build TM5 in each GPCR model. . ϕ_{ρ} , ψ_{ρ} and ω_{ρ} corespond to the middle Pro residue. Adapted from Sankararamakrishnan *et al.* (1991).

AA	χ_1	χ_2	χ_3	χ_4	χ_5	χ_{21}	RRS
Arg	-69	175	174	-75	177		3
Asn	-70	-49					3
Asp	-69	-3					3
Cys	176						2
Gln	-72	174	2				3
Glu	-172	-179	159				3
His	-67	74	180				2
Ile	-60					172	1
Leu	-68	179					2
Lys	-70	-177	176	-177			2
Met	-172	-177	67				2
Phe	-64	-91					1
Pro	-4						1
Ser	-65						2
Thr	63						2
Trp	-160	-104					1
Tyr	-67	-88					1
Val	172						1

Table 9.3.2; Preferred conformers of side-chains Adapted from Sutcliffe *et al.* (1987). RRS = relative reliability scale: 1 = one highly preferred conformer; 2 = two or three highly preferred conformers; 3 = more than three highly preferred conformers, or no highly preferred conformers; AA = amino acid type. χ_1: all except for Thr are g^+ or *trans*. These are expected due to steric hindrance of the C and N for g^- (Thr is able to bend backwards and hydrogen bond to the mainchain); where ± 180 is *trans*, -60^0 is g^+, 60^0 is g^-. The χ_2 value of Pro was not considered in Sutcliffe's study. Also, the value of -4^0 for χ_1 for Pro is not in agreement with the study performed by Milner-White *et al.* (1992) which showed that χ_1 values for Pro are approximately -21^0 (*puckered-up*) or $+22^0$ (*puckered-down*). In this modelling study the χ_1 and χ_2 values used to model every Pro residue (except for TM5 - see table 9.3.1) were: $+18.7^0$ and -14.0^0 respectively (the default values of SYBYL). Following energy minimisation the values typically converged to approx. $+20^0$ and -32^0 for χ_1 and χ_2 respectively.

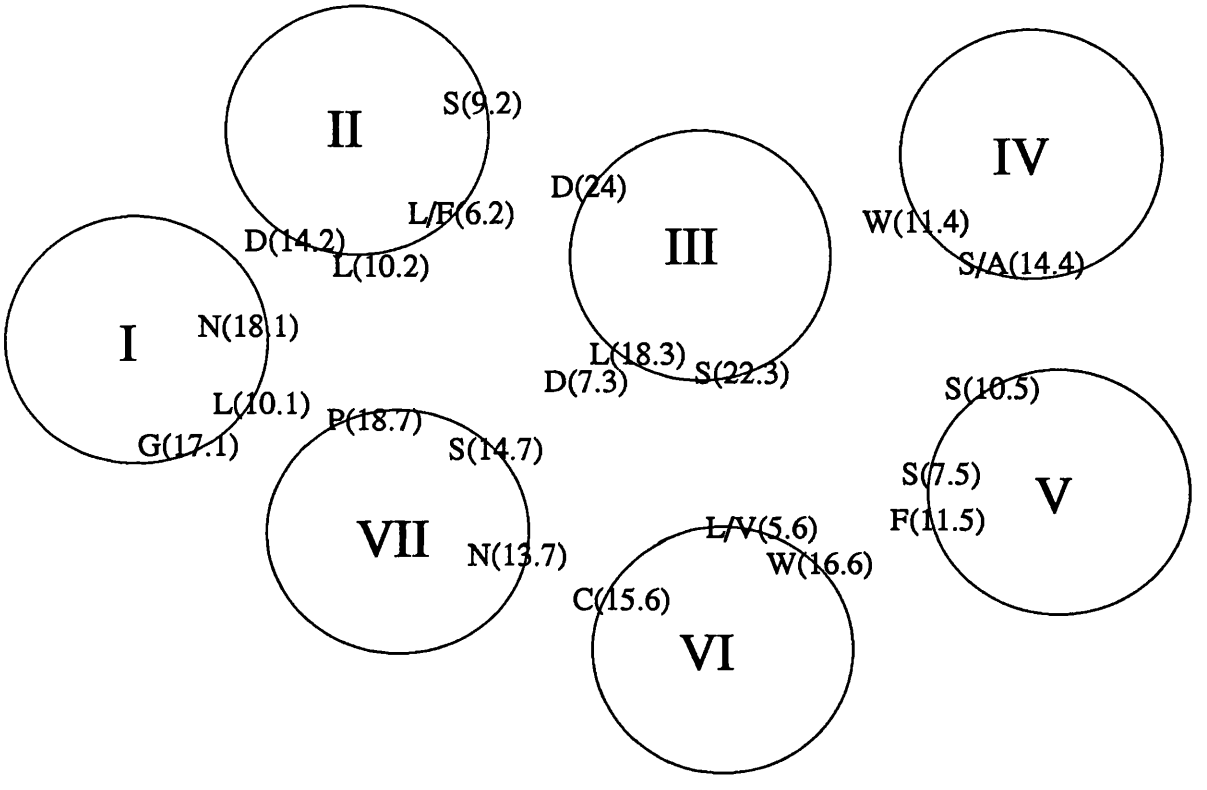
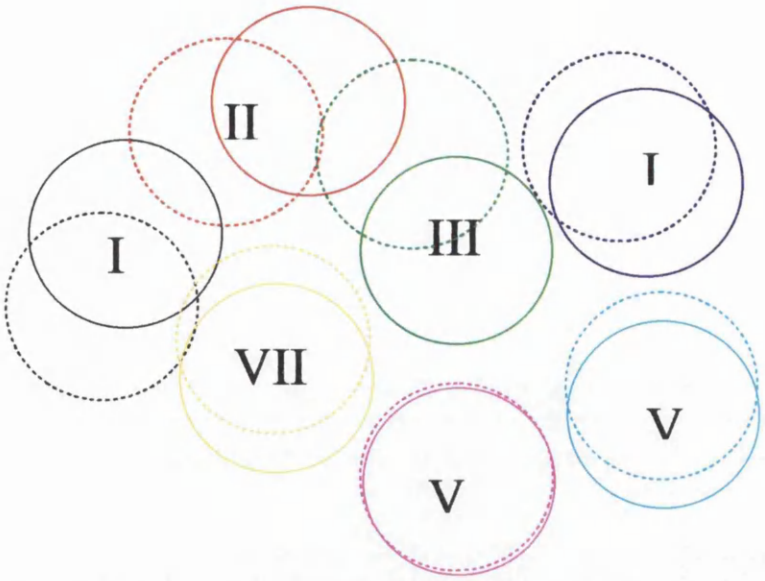


Figure 9.3.3; Schematic representation of helix phase of the dopamine family of receptors viewed from the intracellular surface. Modelled on the work of Joyce Baldwin (1993). Connectivity is clockwise. The figure does not attempt to depict probable helix tilts - see figure 9.3.4.

(a)



(b)

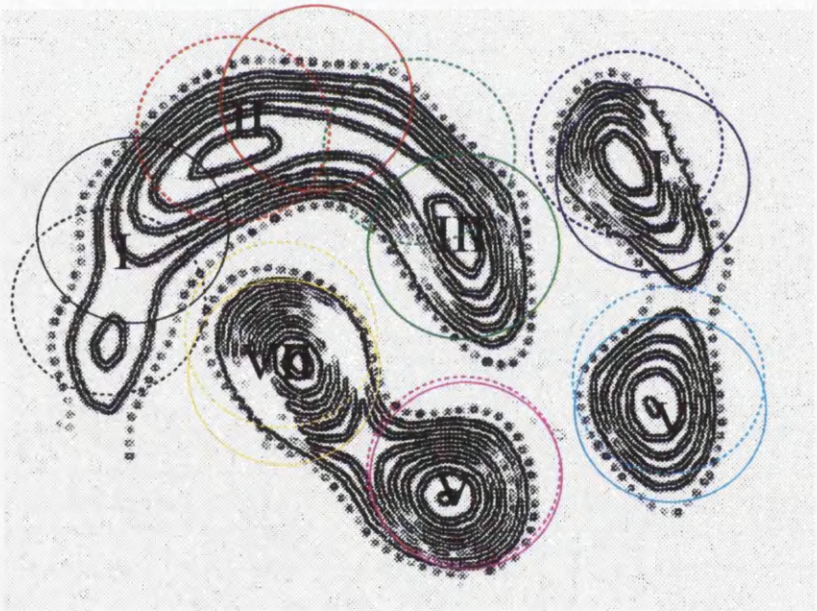


Figure 9.3.4; (a) schematic representation of probable TM helix tilts (used to guide modelling of D_2 and D_1 TM regions) as viewed from the intracellular surface based on the probable arrangement of the helices in GPCRs (Baldwin, 1993). Solid lines indicate intracellular end of TM helices and dotted lines extracellular ends; (b) Projection map of rH (Gebhard *et al.*, 1993) superimposed on modelled TM region.

9.4 Results and Discussion

In reviewing and discussing the results of this modelling exercise it is worthwhile to remind ourselves - in case we need reminding - the limitations inherent in GPCR modelling. Mike Singer (1994) noted that GPCR modelers have a tendency to use phraseology such as: "The GPCR model also demonstrates that ....." Singer commented: "I think we GPCR modelers should be more tentative about the implications of our modelling results. We should not make a practice of saying that the model *demonstrates* a phenomenon; thus far no GPCR model is that reliable. The models may suggest theories, even lend support to them, but in the words of Hibert *et al.* (1993), let's remember that *ce n'est pas un GPCR.*" Some healthy scepticism appears appropriate.

9.4.1 A Physically Plausible Structure?

The case for using a prokaryotic proton pump (bRh) as a template for the construction of dopamine GPCR models can be questioned since bRh lacks sequence homology (chapter 5) with GPCRs. Also, bRh is not coupled to G proteins. In contrast, Rh possesses significant sequence homology with each member of the dopamine family of GPCRs and is also coupled to a G protein. The suggestion that current GPCR models based on bRh are intrinsically flawed has been taken up by Hoflack *et al.,* (1994) who attempted to defend their use of bRh as structural template to model a range of GPCRs - including members of the dopamine family (Trumpp-Kallmeyer *et al.,* 1992). Maloney-Huss and Lybrand (1992) have suggested that a more *de-novo* modelling approach is applicable to modelling physically plausible structures of GPCRs. It is not surprising therefore that the TM regions of the GPCRs modelled in this study do appear different from the bRh transmembrane region (see figure 9.4.1).

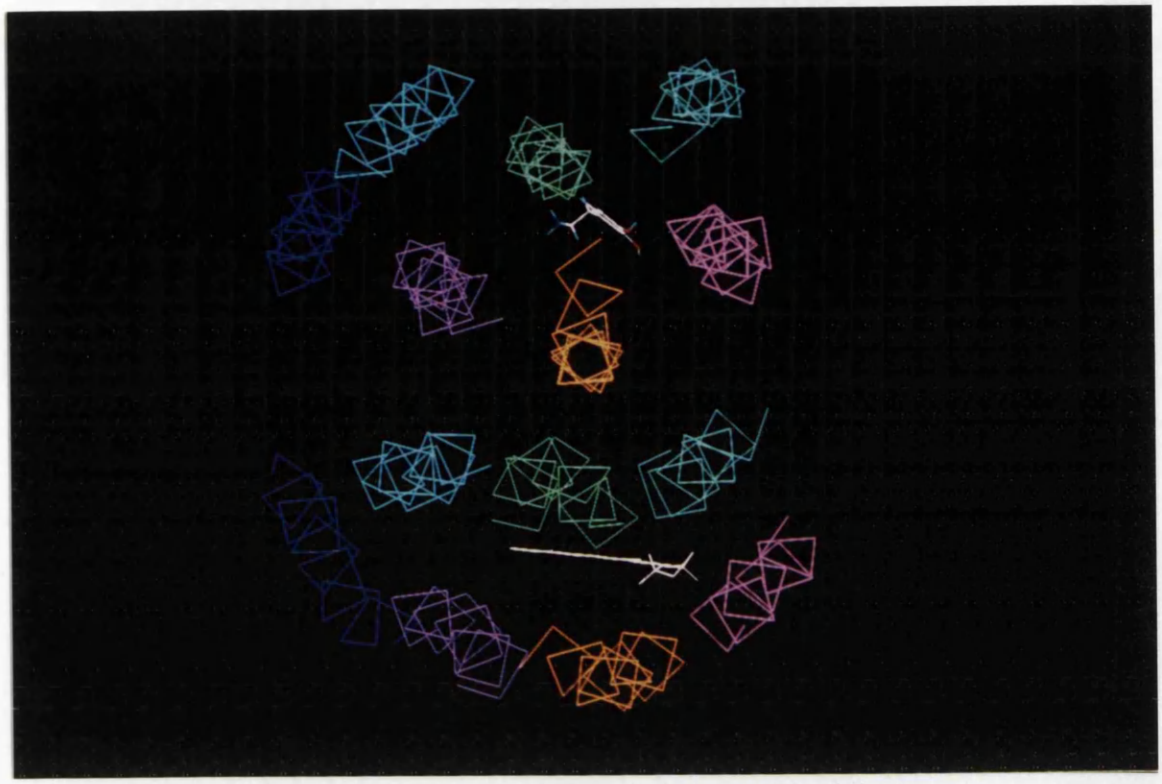


Figure 9.4.1 Comparison of C^{α} plots of the transmembrane regions of D_2 (top) and bacteriorhodopsin (bottom) viewed from the intracellular side of the membrane. The TM helices numbering is clockwise from TM1 (blue), TM2 (blue-green), TM3 (green), TM4 (cyan), TM5 (magenta), TM6 (orange) and TM7 (purple). The TM region of D_2 clearly complements the 2D projection map of rhodopsin (figure 9.3.4) whereas the TM region of bRh clearly does not.

Ramachandran plots of the dopamine receptors and the muscarinic M1 receptor (table and figure 9.4.1.1) clearly show that each of the models have very favourable phi/psi angles. Comparison with the recently released model of Donnelly *et al.* (1994) provides a useful bench-mark to judge the models generated for this Ph.D (figure 9.4.1.1). The models generated in this study were energy minimised to between -1,653kcals/mol and -1860kcals/mol (table 9.4.1.2). The point being to ensure that the models were stable and in that sense physically plausible structures.

Of particular interest is the distorted geometry summary produced by Procheck (figure 9.4.1.3; also refer to appendix 5). It is clear that each of the models produced in this study have some distorted geometry in some of the planar groups. Given more time, the author

would use this data to go back to the models and attempt to alleviate the abnormal stresses that are causing the distortion in some of the planar groups. However, it is quite clear that the models generated in this study compare favourably with the Donnelly β_2-AR model in terms of distorted geometry - particularly with regard to the distorted main chain bond lengths and main chain bond angles (table 9.4.1.3). Procheck did not find any distorted main-chain bond angles or main chain bond lengths in any of the models generated in this study.

	D_1		D_2		D_3		D_4		D_5		M1		β_2	
	No	%	No	%	No	%	No	%	No	%	No	%	No	%
Residues in most favoured regions [A,B,L]	157	98.1	154	96.9	154	98.1	150	98.7	157	98.7	156	99.4	178	92.7
Residues in additional allowed regions [a,b,l,p]	3	1.9	3	1.9	3	1.9	2	1.3	2	1.3	1	0.6	12	6.2
Residues in generously allowed regions [~a,~b,~l,~p]	0	0.0	1	0.6	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Residues in disallowed regions	0	0.0	1	0.6	0	0.0	0	0.0	0	0.0	0	0.0	2	1.0
Number of non-glycine and non-proline residues	160	100.0	159	100.0	157	100.0	152	100.0	159	100.0	157	100.0	192	100.0
Number of end residues (excl. Gly and Pro)	14		14		14		14		14		14		14	
Number of glycine residues	3		4		6		8		4		7		10	
Number of Proline Residues	5		5		5		8		5		4		5	
Total number of residues	182		182		182		182		182		182		221	

Table 9.4.1.1 Ramachandran Plot results of Phi against Psi in tabular form for D_1 through to D_5 dopamine receptors and muscarinic M1 GPCR. β_2-adrenergic receptor is the model of Donnelly *et al.* (1994). Statistics obtained using Procheck (Laskowski *et al.*, 1993). See figure 9.4.1.1 for explanation of meaning of A, B, L, a, b, l, p, ~a, ~b, ~l, ~p. Statistical analysis was based on 118 structures of at least 2.0 Angstroms and R-factor greater than 20%. A good quality model would be expected to have over 90% of phi/psi angles in the most favoured regions.

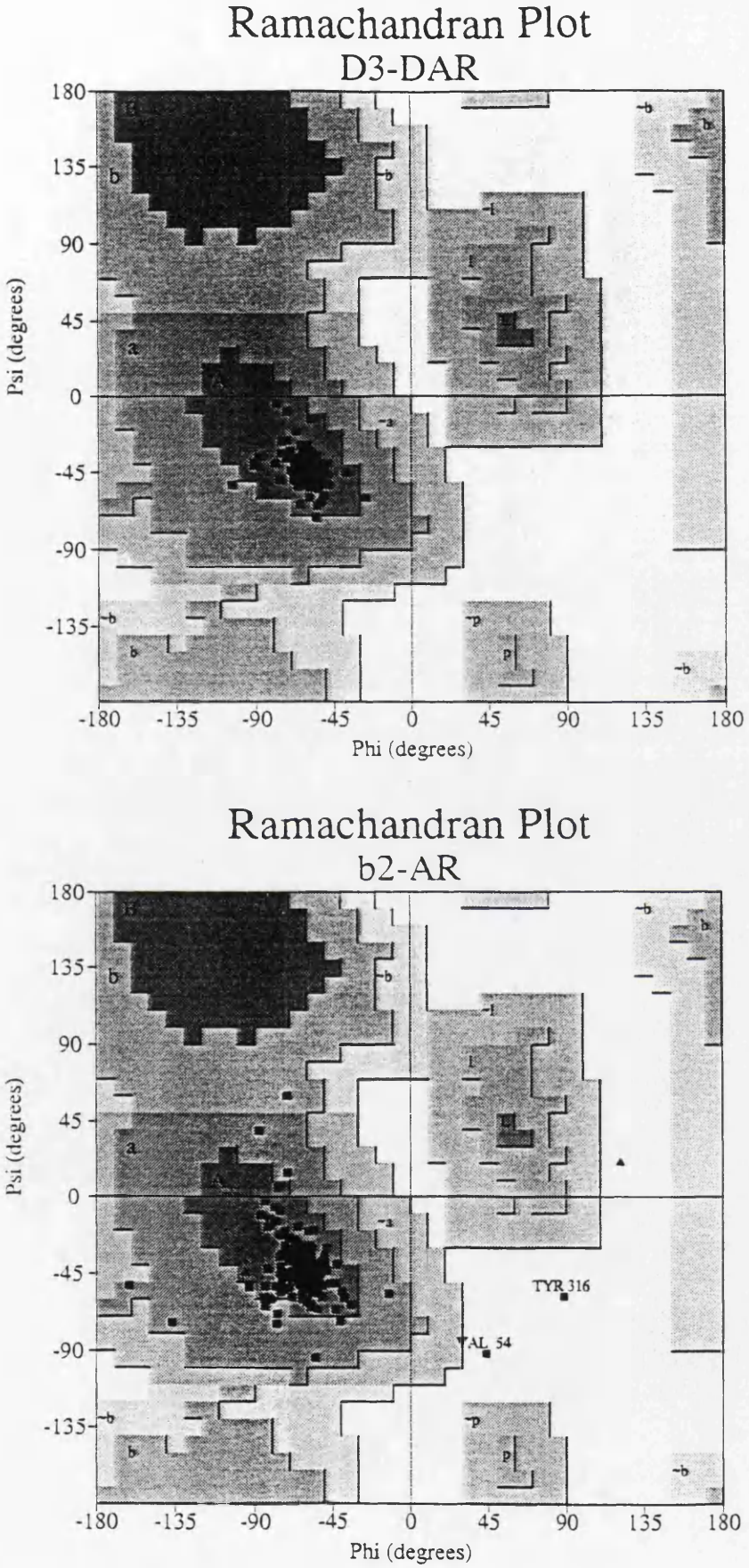
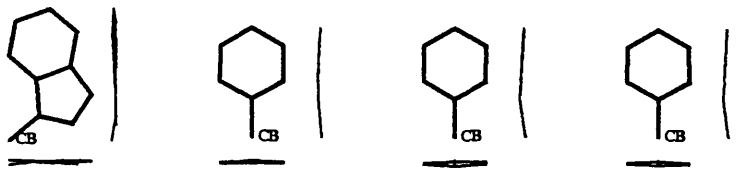


Figure 9.4.1.1 Ramachandran Plots. Top: D_3 dopamine receptor (D3-DAR) and for comparison (bottom) Donnelly's model of β_2-adrenergic receptor (Donnelly *et al.*, 1994). Regions A and a correspond to residues in alpha helix conformation. Residues in disallowed regions are labelled.

Distorted geometry
D1-DAR

Planar groups

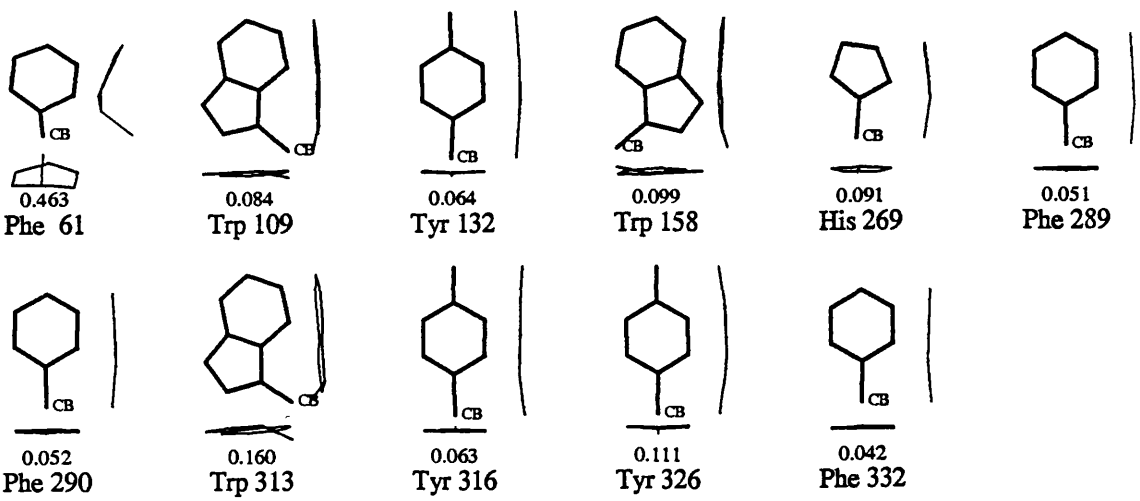


0.061 B Trp 24 | 0.046 F Phe 19 | 0.075 F Phe 20 | 0.065 G Phe 3

Sidechains with RMS dist. from planarity > 0.04A for rings, or > 0.03A otherwise. Value shown is RMS dis

Distorted geometry
b2-AR

Planar groups



0.463 Phe 61 | 0.084 Trp 109 | 0.064 Tyr 132 | 0.099 Trp 158 | 0.091 His 269 | 0.051 Phe 289

0.052 Phe 290 | 0.160 Trp 313 | 0.063 Tyr 316 | 0.111 Tyr 326 | 0.042 Phe 332

Sidechains with RMS dist. from planarity > 0.04A for rings, or > 0.03A otherwise. Value shown is RMS dist.

Figure 9.4.1.2 Top: distorted geometry in D_1 dopamine receptor (DAR). Bottom: β_2-adrenergic receptor (AR) - Donnelly *et al.*, 1994. In addition to the distorted planar groups that Procheck (Laskowski *et al.*, 1993) found in β_2-AR model there were numerous distorted main-chain bond lengths and some distorted main-chain bond angles (table 9.4.1.3). For full output of distorted geometry see appendix 5.

	D_1 kcals/mol	D_2 kcals/mol	D_3 kcals/mol	D_4 kcals/mol	D_5 kcals/mol	M1 kcals/mol
Bond Stretching Energy	21.1	21.3	20.2	21.5	20.9	20.1
Angle Bending Energy	110.6	323.7	107.2	123.0	113.5	95.7
Torsional Energy	145.9	139.5	154.5	167.0	154.1	134.9
Improper Torsional Energy	24.3	26.9	20.4	22.3	22.7	19.9
1-4 van der Waals Energy	353.8	357.6	350.1	313.5	353.1	336.6
van der Waals Energy	-1364.6	-1390.8	-1336.9	-1244.6	-1360.3	-1335.3
1-4 Electrostatic Energy	1388.1	1373.79	1319.9	1304.5	1353.9	1322.8
Electrostatic Energy	-2452.6	-2412.8	-2366.3	-2350.2	-2424.8	-2356.2
H-Bond Energy	-86.5	-92.3	-90.3	-80.1	-88.4	-82.7
Total Energy	-1859.8	-1653.0	-1820.6	-1723.1	-1855.3	-1844.2

Table 9.4.1.2 Final energy values for each model.

	D_1	D_2	D_3	D_4	D_5	M1	β_2
Main-chain bond lengths (N - CA)	0	0	0	0	0	0	11
Main-chain bond lengths (CA - C)	0	0	0	0	0	0	109
Main-chain bond angles (CA - C - N)	0	0	0	0	0	0	2
Main-chain bond angles (CB - C - N)	0	0	0	0	0	0	5
Main-chain bond angles (N - CA - CB)	0	0	0	0	0	0	1
Planar groups - rings	4	8	6	3	7	5	11
Planar groups - other	0	0	0	0	0	0	0

Table 9.4.1.3 Distorted geometry as determined by Procheck (Laskowski *et al.*, 1993). β_2-adrenergic receptor (Donnelly *et al.*, 1994) clearly has a considerable amount of distorted geometry especially in terms of distorted main-chain bond lengths. Procheck flags bonds differing by > 0.005Å from ideal small-molecule values and bond angels differing by > 10.0 degrees from small-molecule values. Procheck flags side-chains with RMS distance from planarity > 0.04Å for rings, or >0.03Å otherwise.

9.4.2 Helix Kinking Due To Middle Proline Residues

Kinking of TM helices was definitively demonstrated by Henderson *et al.* (1990) in their model of the structure of bacteriorhodopsin based on high resolution electron cryo-microscopy. However, their model was not of sufficient resolution to provide definitive torsional backbone angles. Figure 9.4.2 compares the kinks caused by middle Pro residues with a regular TM helix. It is clear that TM5 and TM7 are kinked. TM6 (not shown) is also kinked since it has a middle Pro residue in each member of the dopamine family of GPCRs.

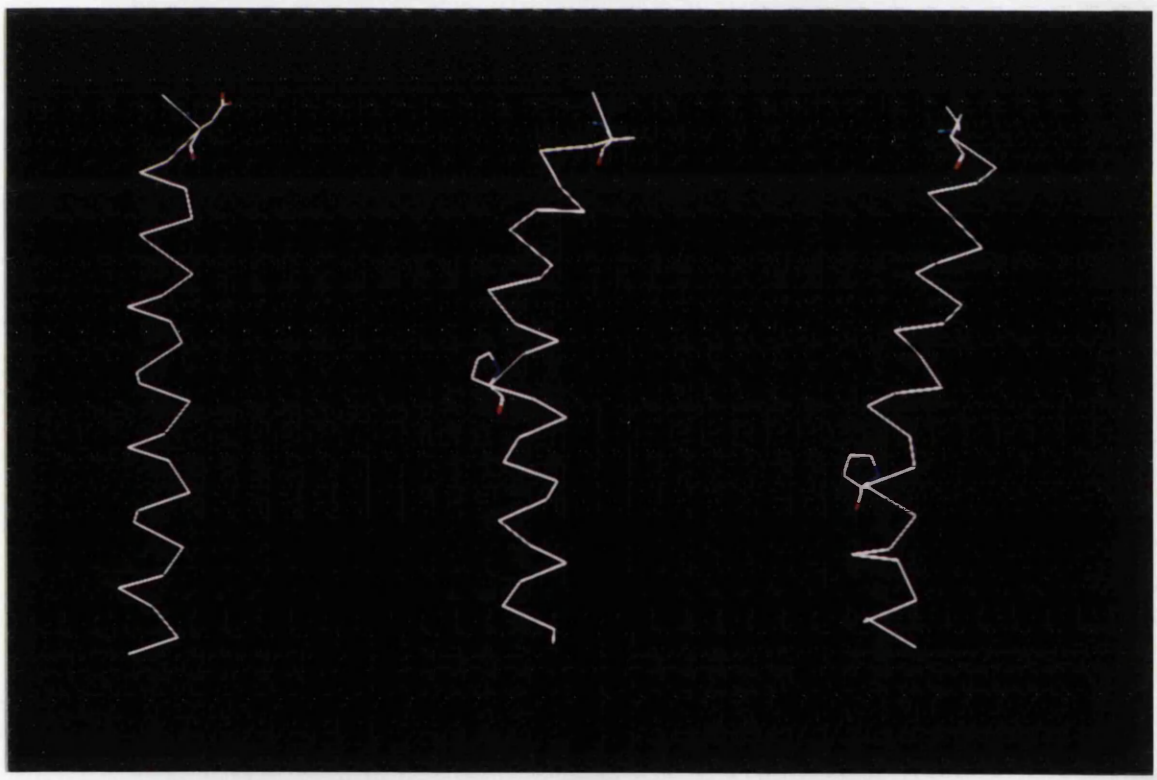


Figure 9.4.2; the effect of a middle Pro residue in TM helices extracted from D_2. Far left: TM3 which lacks a middle Pro residue and was modelled solely using ϕ/ψ angles of: -59^0 and -44^0 respectively for a helix in a hydrophobic environment (Blundell *et al.*, 1983). Middle: TM5 with a middle Pro which was modelled using the main-chain torsion angles suggested by Sankararamakrishnan *et al.*, 1991. Right: TM7 has a Pro residue and was modelled using the *Trans-1* configuration suggested by Polinsky *et al.*, 1992. The first residue in each TM helix is also shown.

9.4.3 Interhelical Hydrogen Bonding

Each of the dopamine receptor models (also the model of M_1) were examined for interhelical hydrogen bonding by using hydrogen bond display option in SYBYL. The object here being to examine the nature of possible hydrogen bonding in GPCRs - see table 9.4.3.1 Figure 9.4.3.1 illustrates the interhelical hydrogen bonding found in the models of D_3 and M_1. Of particular interest are interhelical hydrogen bonding involving conserved residues which occupy middle positions in TM helices (i.e. are at least 5 residues from the amino or carboxy termini). Such residues are less likely to be hydrogen bonded to the polar end groups of the surrounding lipid molecules of the cell membrane.

Asp_{214} is conserved in catechol amine binding GPCRs and other classes of GPCRs; Probst *et al.* (1992) has noted that Asp_{214} is 98% conserved in the superfamily of GPCRs. It is believed to play a key role in the binding of agonist ligands. In the model of D_2, Asp_{214} forms two hydrogen bonds - one with conserved Asn_{118} and the other with conserved Ser_{714}. In the remaining models Asp_{214} forms a hydrogen bond with either Asn_{118} or Ser_{714}. The model of β_2-adrenergic receptor built by Donnelly *et al.* (1994) was obtained and briefly examined using SETOR (Evans, 1993) and the interhelical hydrogen bonding pattern displayed; a hydrogen bond was found in the model between Asn-51 (equivalent to Asn_{118}) and Asp-79 (Asp_{214}). There is also a strongly conserved Asn residue in TM7 (Asn_{717}) which is conserved in 95% of all GPCRs (Probst *et al.*, 1992). On the basis of mutagenesis studies, Zhou *et al.* (1993) have suggested that the equivalent Asn residue is involved in hydrogen bonding to the equivalent Asp residue in the gonadotropin-releasing hormone GPCR. Hence, a hydrogen bonding network involving both conserved Asn residues (Asn_{118} and Asn_{717}) and Asp_{214} may play an important role in the functioning of GPCRs. It is likely that the phase of TM7 in the models generated here needs to be adjusted so that Asn_{717} is in a position that renders it more amenable to hydrogen bonding to Asp_{214}. Given that the helix phase used for orientating TM7 is based on the conclusions of Baldwin (1993), the findings of Zhou *et al.* (1993) suggest that Baldwin's proposed orientation for TM7 is wrong. In the Donnelly *et al.* (1994) model TM7 is so orientated that Asn_{717} is hydrogen bonded to Asn_{623}.

Another interesting interhelical hydrogen bonds which crops up in most of the models generated in this study is the hydrogen bond between the conserved Trp_{411} and the back-bone carbonyl group of Ile_{315} or Val_{315} for the muscarinic M1 receptor. Clearly the large size of Trp renders it amenable to hydrogen bonding to a carbonyl group on an adjacent TM helix. This hydrogen bond is lacking in the Donnelley *et al.* (1994) model of the β_2-adrenergic receptor.

Of particular novelty is the hydrogen bonding pattern between TM1 and TM2 in the model of D_5 - figure 9.4.3.2. The main chain carbonyl oxygen of Leu_{210} is hydrogen bonded to the side-chain of Asn_{118} (HD22) and the main chain amine of Asp_{214}. The side-chain of Asp_{214} (OD2) is in turn hydrogen bonded to HD22 of Asn_{118}. The side-chain geometry of Asp_{214} also renders OD2 amenable to hydrogen bonding to its main chain NH group. This produces a twisted parallogram form of hydrogen bonding.

GPCR TYPE	Residue	Hydrogen bonded to:
D_1	Asp_{214}	Ser_{714}
D_1	Asp_{324}	Tyr_{721}
D_2	Asp_{214}	Asn_{118} , Ser_{714}
D_2	Thr_{305}	Asn_{426}
D_2	Trp_{411}	Carbonyl of Ile_{315}
D_2	Lys_{524}	Tyr_{326}
D_3	Arg_{704}	Carbonyl of Tyr_{103}
D_3	Asp_{214}	Asn_{118}
D_3	Asp_{324}	Tyr_{721}
D_3	Tyr_{326}	Arg_{524}
D_3	Trp_{411}	Carbonyl of Ile_{315}
D_4	Ser_{124}	Pro_{718}
D_4	Asp_{214}	Asn_{118}
D_4	Asp_{324}	Tyr_{721}
D_4	Trp_{411}	Carbonyl of Ile_{315}
D_5	Asn_{118}	Asp_{214}, carbonyl of Leu_{210}
D_5	Asp_{324}	Tyr_{721}
D_5	Ser_{606}	Tyr_{522}
M1	Asp_{214}	Asn_{118}
M1	Ser_{221}	Tyr_{707}
M1	Asp_{324}	Tyr_{721}
M1	Trp_{411}	Carbonyl of Val_{315}

Table 9.4.3.1 Interhelical hydrogen bonding in the models of the dopamine (DA) G protein coupled receptors (GRCRs): D_1 through to D_5 and the muscarinic M1 receptor. First digit of subscript number refers to transmembrane (TM) helix number and the remaining two digits represents residue position in the TM helix, for example: Asp_{214} is the 14th residue in TM2. Unless otherwise stated, all interhelical hydrogen bonding is side-chain to side-chain.

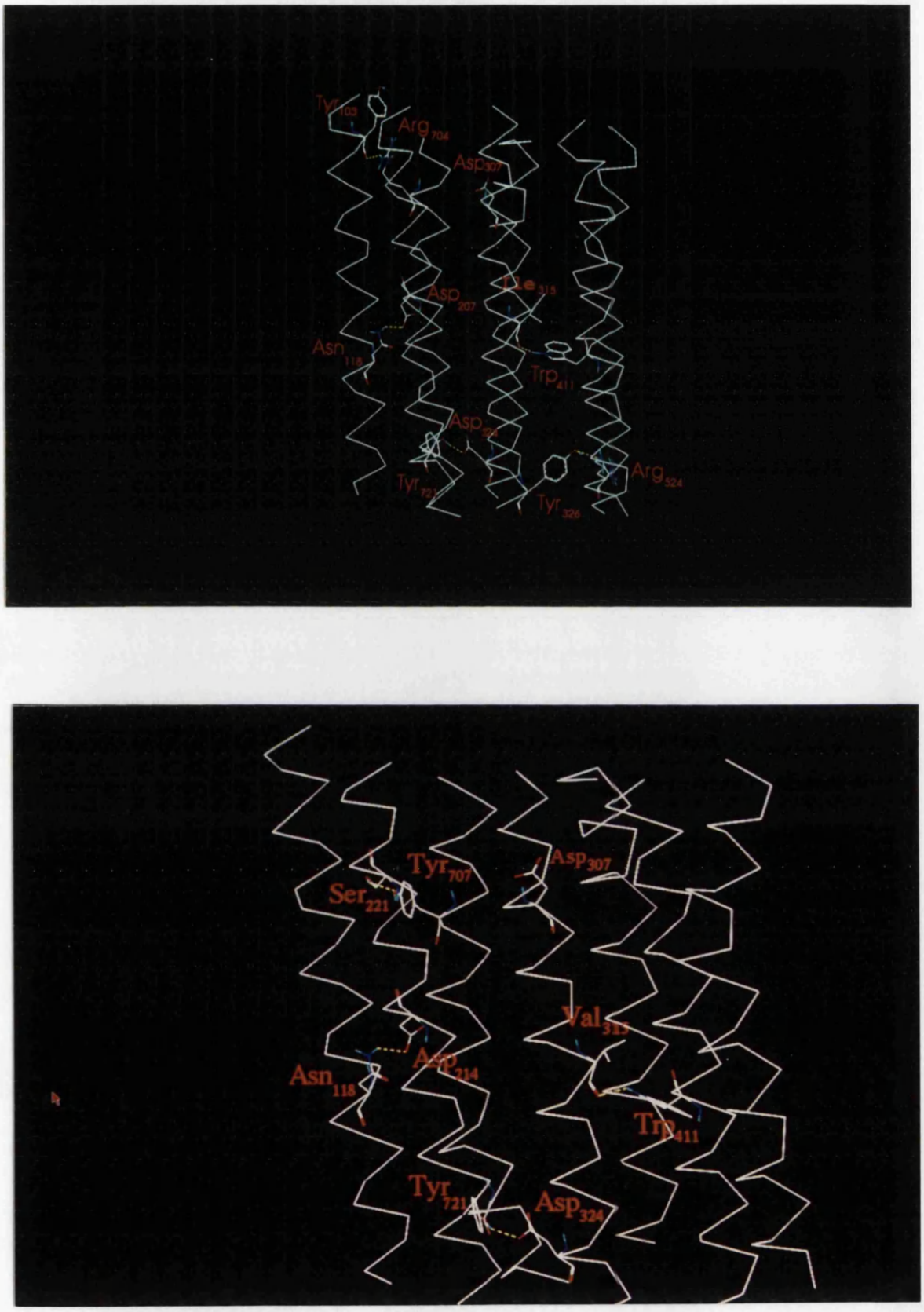


Figure 9.4.3.1 Interhelical hydrogen bonding found in the models of D_3 (top) and M_1 (bottom).

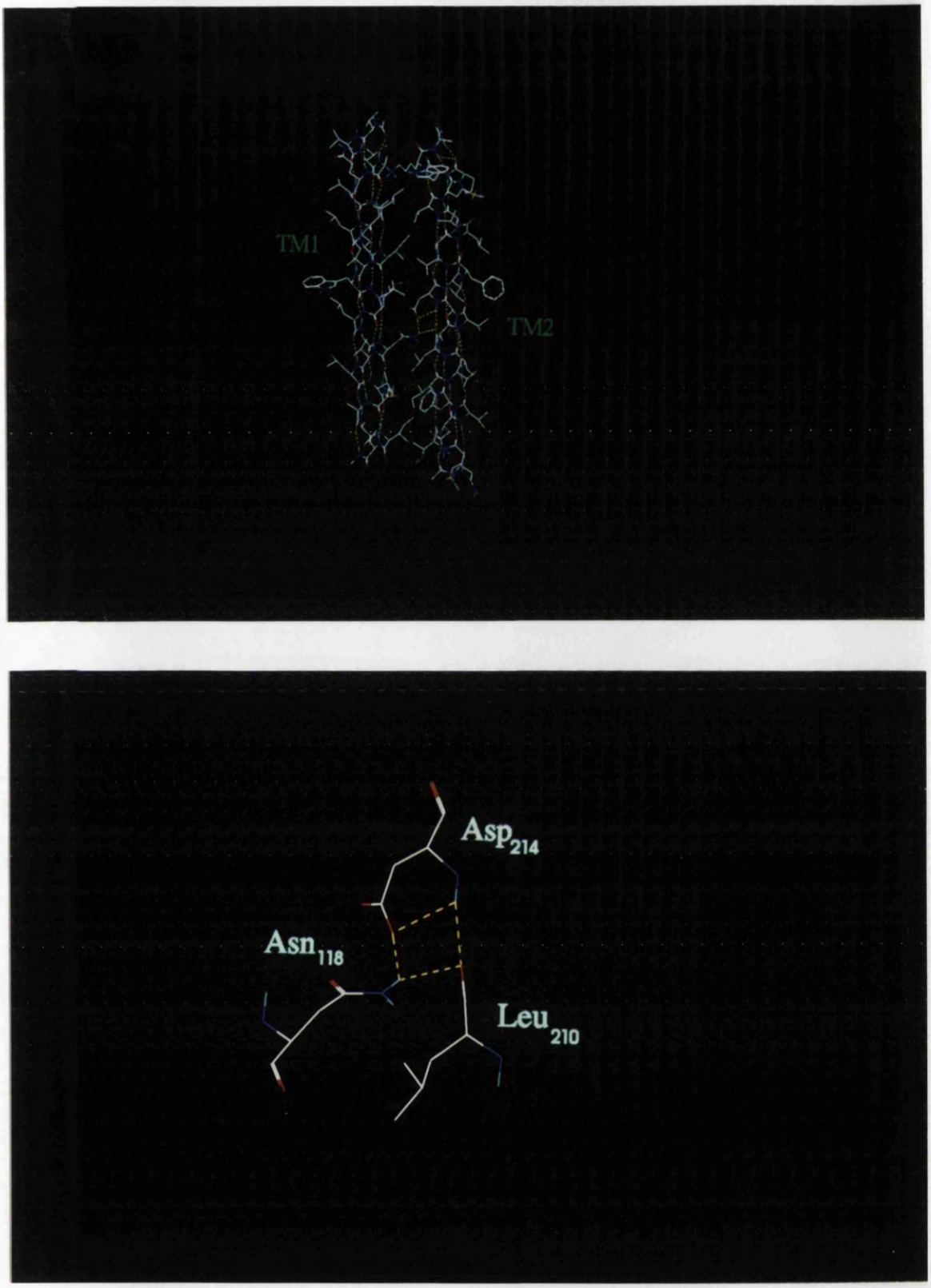


Figure 9.4.3.2 Top: the unusual aesthetic feature of the interhelical hydrogen bonding between TM1 and TM2 in the model of D_5 - a twisted parallogram is clearly discernable. Bottom: Close up of the interhelical bonding between TM1 and TM2.

9.4.4 Intrahelical Hydrogen Bonding Pattern

A further outcome leading from the presence of a middle Pro residue is that the intra-helical hydrogen bonding pattern is disrupted. The presence of a cyclic side chain ~~not only~~ prevents the formation of a hydrogen bond with the preceding turn of the helix, because of the absence of the amide H atom (reviewed by Piela *et al.*, 1987). This means hydrogen bonding with the carbonyl oxygen atom of residue *i-4* (where *i* is the position of the Pro residue) is missing - see figure 9.4.4.. The *i-4* residue is Ser_{510} which also happens to be involved in hydrogen bonding to the agonist ligand. The lack of a back-bone hydrogen bond between conserved Ser_{510} and Pro_{514} in conjunction with the resultant kinking of the TM5 are features which probably work together to aid agonist ligand binding.

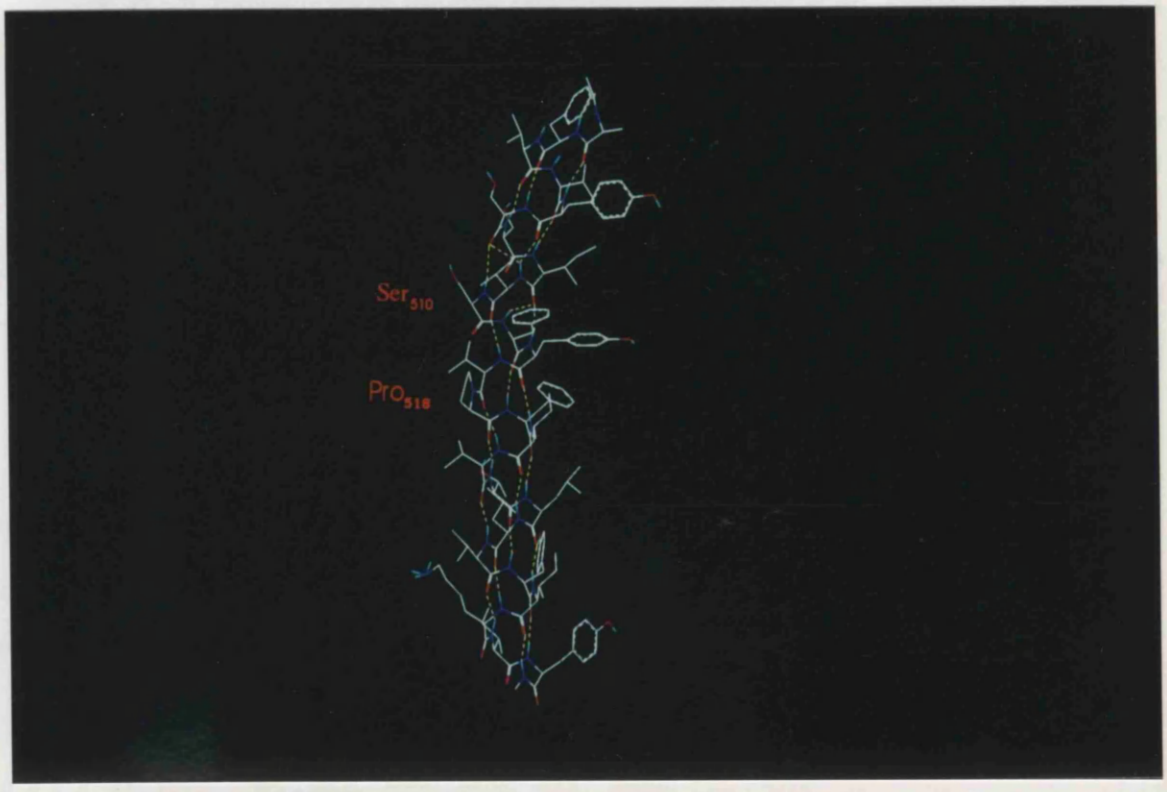


Figure 9.4.4 The effect of the conserved middle Pro residue in TM5 of D_2. The Pro_{514} residue (at position *i)* is lacking a back-bone N-H group and so can not hydrogen bond to the *i-4* residue (Ser_{510}).

9.4.5 Ligand Binding To Dopamine GPCRs

The natural agonist ligand dopamine incorporates a quaternary ammonium group and a catechol moiety which in turn incorporates an aromatic moiety. Hence, this suggests that four main interactions are possible:

- hydrogen bonding between the hydroxyl groups of the catechol ring and suitable side-chains such as Ser and Thr.

- reinforced ionic bonding between the positively charged ammonium group and appropriate negatively charged side residues such as Asp and Glu.

- σ - π interactions of the type described by Verdonk *et al.*, (1993) who described charged nitrogen-aromatic interactions in ligand-receptor binding. Where the "charged nitrogen" is the positively charged quaternary ammonium group of the catechol amine.

- π - σ interactions of the type described by Burley and Petsko (1989) such as interactions between the $\delta(-)$ π-electron cloud of the aromatic moiety of dopamine and $\delta(+)$ amino groups in such residues as e.g. Lys and Asn.

The last two interactions are beyond the scope of this modelling exercise to demonstrate. However, the likely interaction of the natural agonist in terms of hydrogen bonding and ionic bonding can be considered.

9.4.5.1 Hydrogen Bonding

The multiple sequence file (MSF - appendix 1) reveals that two Ser residues are conserved in the catechol amine binding GPCRs in putative transmembrane helix 5 (TM5). Using Baldwin's (1993) transmembrane helix assignments (see figure 9.3.2) the positions of

adrenergic receptor have been shown experimentally to be play a critical role in the binding and activation of this receptor by catecholamine agonists. Strader *et al.* (1989a) showed that replacement of either Ser residue with an Ala residue resulted in a reduction in the affinity and efficacy of catecholamine agonists at the receptor, with no alteration in antagonist binding. In the docking of the dopamine ligand to D_2 considerable care was therefore taken to ensure that the natural ligand was orientated in such a manner that its catechol moiety could form hydrogen bonds with both conserved Ser_{507} and Ser_{510}. However, after several attempts it became evident that only the *p*-hydroxyl group of the catechol moiety could hydrogen bond with Ser_{510}; see figures 9.4.5.1.1 and 9.4.5.1.2. The *m*-hydroxyl group simply did not adopt the right orientation to bond Ser_{507}. This observation is particularly interesting as Trumpp-Kallmeyer *et al.* (1992) observed in their model of the dopamine ligand/D_2 complex that the *m*-hydroxyl group hydrogen bonded particularly well with Ser_{507} (Ser_{505} in their personal assignment of amino-acids to TM helices). The hydrogen bond between the *p*-hydroxyl group and Ser_{510} (Ser_{507} in their model of TM5) was weak.

In the Trumpp-Kallmeyer *et al.* (1992) the backbone ϕ/ψ torsion angles used to model each TM helix was: -59^0 and -44^0 respectively. No real attempt was made to use backbone torsion angles to model TM helices with middle Pro residues. In contrast, the back-bone torsion angles suggested by Sankararamakrishnan *et al.* (1991) for modelling a kinked helix caused by a middle Pro residue was used here to model TM5, which has a middle Pro residue. Hence in the bent configuration Ser_{510} is available to hydrogen bond to the *p*-hydroxyl group and in the straight configuration Ser_{507} is available to hydrogen bond to the *m*-hydroxyl group. These different observations add credence to the theory that TM5 is oscillating between a largely straight secondary structure and a kinked one.

Strader *et al.* (1989a) carried out observations on wild-type and mutant β_2 adrenergic receptors and concluded that Ser^{204} (Ser_{507} using the numbering scheme adopted here) hydrogen bonds to the *m*-hydroxyl group of the agonist ligand and Ser^{207} (Ser_{510}) hydrogen bonds to the *p*-hydroxyl group. They also noted that analogs of the natural agonist ligand which lack one of the hydroxyl groups on the aromatic ring function almost as well as the natural agonist. This fits in with the theory that TM5 is oscillating between a largely

straight secondary structure and a kinked one allowing alternating hydrogen bonding and functions as a guiding arm - driving the agonist deeper into the catechol amine GPCR (see chapter 8; figures 8.3.2.1 to 8.3.2.3). Such a mechanism of agonist binding is not in keeping with the classic lock-and-key fit of the agonist to the binding pocket. The probable simultaneous role of helix interface shear mechanism (reviewed in chapter 8) along with the kinking action of the middle Pro residue in TM5 suggests a series of successive stages in keeping with the "zipper" mechanism. It seems likely that kinking of TM6 and TM7 must play a critical role in the final binding of the agonist to Asp_{214}. A similar hypothesis has been put forward by Dahl *et al.*, (1991a) who concluded that D_2 agonist binding involves Asp_{214} and is driven by a "zipper" binding mechanism.

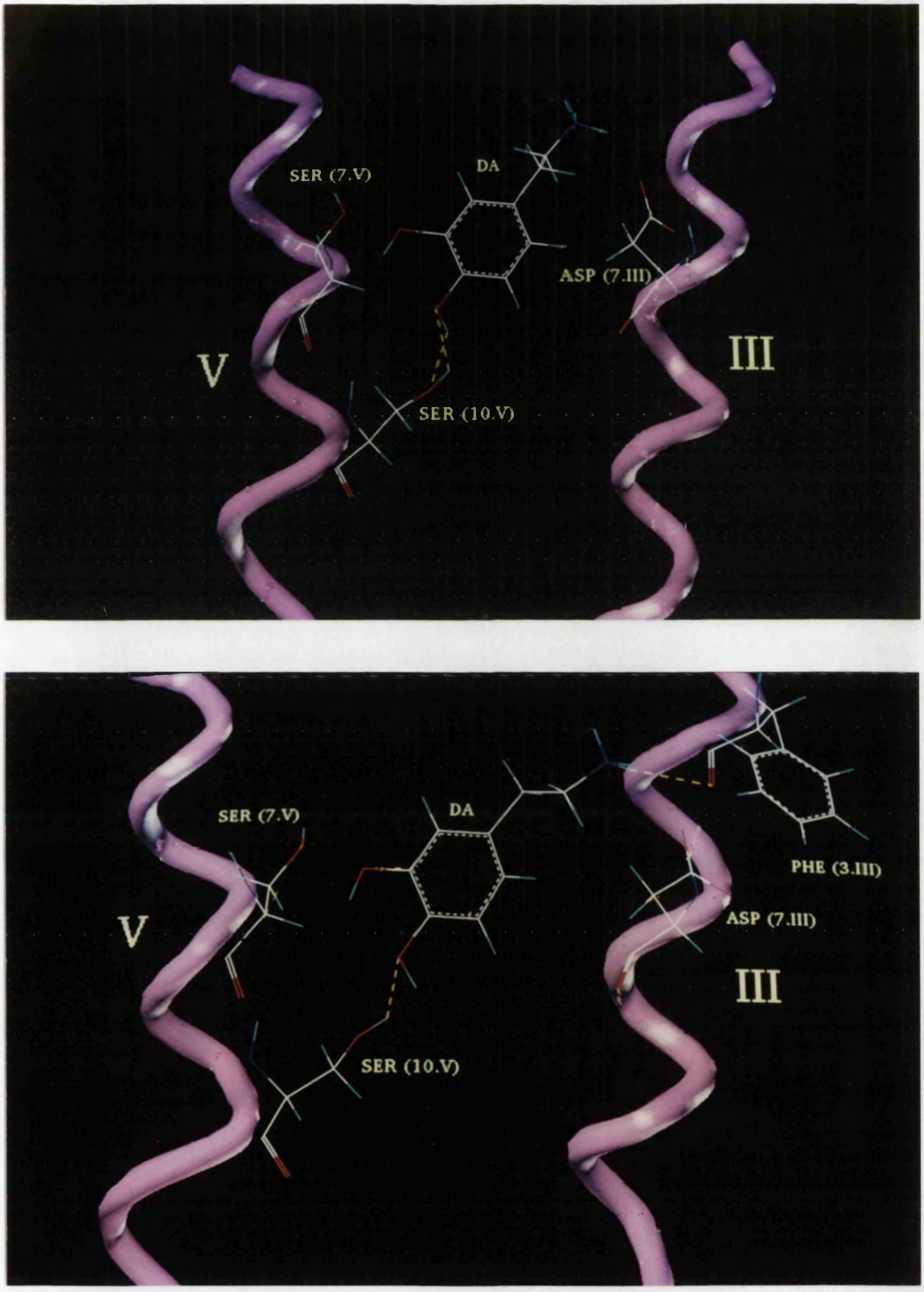


Figure 9.4.5.1.1 Top: hydrogen bonding between *p*-hydroxyl dopamine (DA) and Ser_{510} (D_2 model) after just 20 iterations of energy minimisation (following manual docking to the energy minimised model of D_2). Bottom: single hydrogen bond between *p*-hydroxyl on catechol ring and Ser_{510} after 30 iterations of energy minimisation. It is clear that the DA is moving away from Ser_{510}. No hydrogen bonding is seen between *m*-hydroxyl group and Ser_{507}.

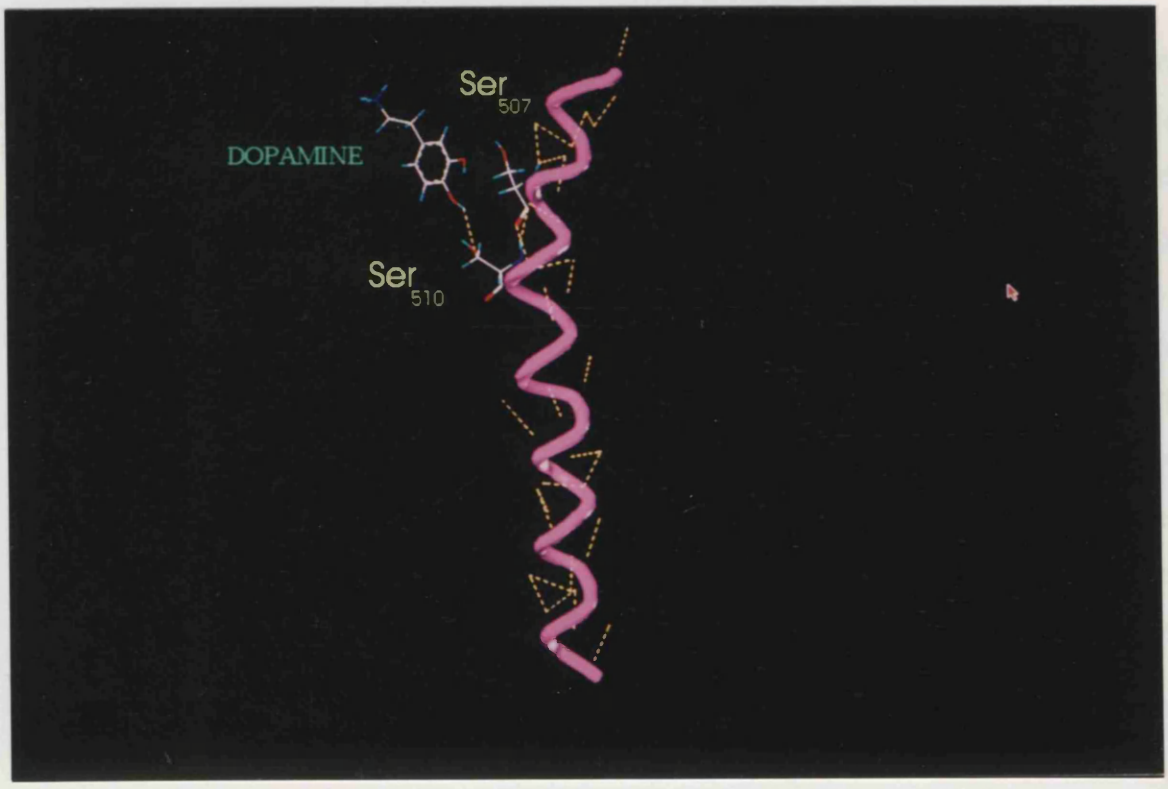


Figure 9.4.5.1.2 To give some idea of scale the dopamine ligand (DA) is shown along with TM5 (both of which have been extracted from manually docked model of the D_2 receptor model after just 30 iterations of additional energy minimisation with mild distance constraints). Again the hydrogen bonding between TM5 and DA is shown.

9.4.5.2 Reinforced Ionic Bonding

Vlijmen and IJzermann (1989) examined the role of reinforced ionic bonding in the binding of ligands to TM3 of β_2-adrenergic receptor. They noted an interaction between what they described as a positively charged nitrogen in the ligand (*S*-PROPRANOLOL) and Asp^{113} (Asp_{307} using the convention adopted here). They declared this interaction was of the type described by Albert (1971, 1979); namely: a reinforced ionic bond. The distance between the nitrogen-oxygen(Asp) was found to be 3.4Å.

Figure 9.4.5.2.1 illustrates the natural agonist in the putative binding pocket of D_2. While it is clear that the quaternary ammonium group of DA is in close proximity to Asp_{307}

(conserved in all members of the catechol amine binding class of GPCRs) it is evident that the hydrogen bond that characterises reinforced ionic interactions is not evident between the "positively" charged nitrogen and one of the oxygen's of Asp_{307}. As a guide to the nature of the likely interaction between the "positive nitrogen" and Asp_{307} the charge distribution around dopamine was calculated using the Geister-Huckel method (chapter 2; figure 2.2.3.1). It is clear that the nitrogen is not the classic "positively charged nitrogen" alluded to in the literature describing the putative binding of the catechol amines. The positive charge of the "nitrogen" is infact spread around the quaternary ammonium group. Hence, declaring the presence of a reinforced ionic bond on the basis of the distance between the "nitrogen" and one of the oxygen's of Asp_{307} is rather suspect. During the very sensitive energy minimisation of the DA agonist in the putative binding pocket the closest distance between Asp_{307} and the quaternary ammonium group was ≈3Å.

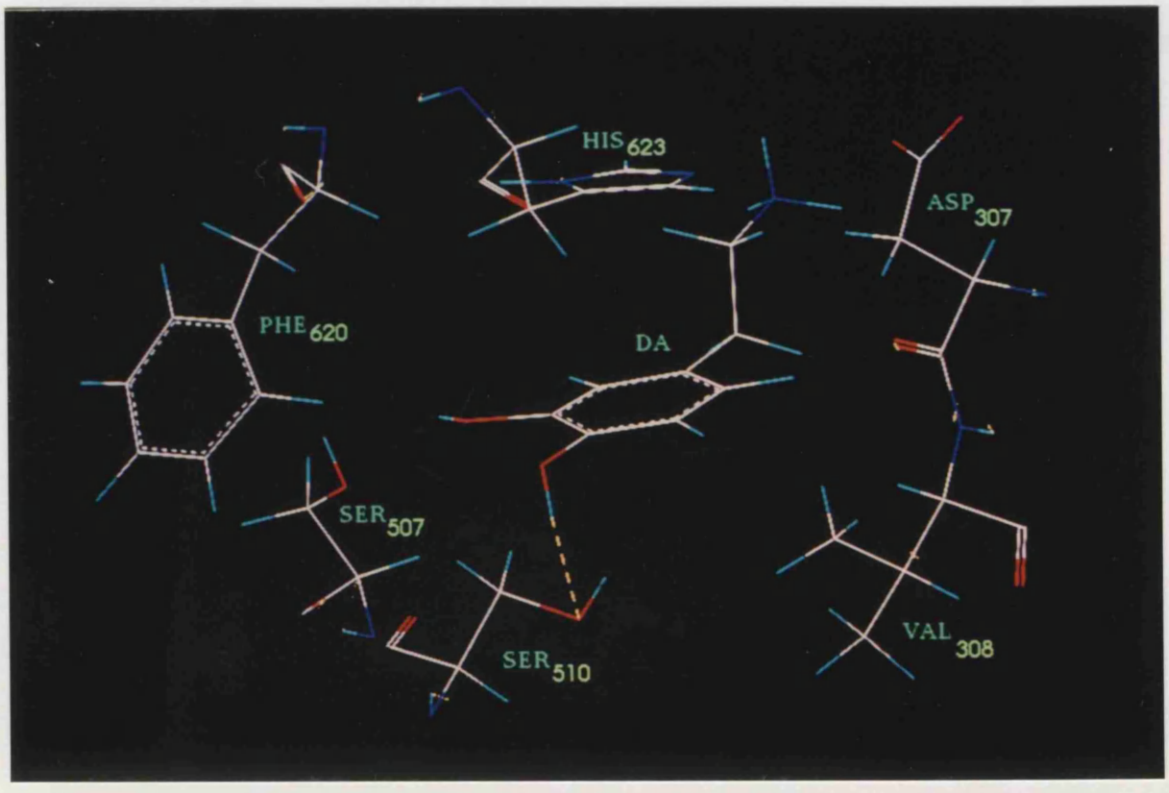


Figure 9.4.5.2.1 Dopamine in the interior of D_2 close to the extracellular surface. This is a 3D representation of Mode 0 as envisaged by Saunders and Findlay (1990) - see figure 3.4.1 on page 36. The kinking action of TM5 will drive the agonist deeper into the receptor enabling interaction with Asp_{214}.

9.4.6 Cavities In D_2

(Nicholls *et al.*, 1993) developed a rapid surfacing and visualisation program: GRASP (Graphical Representation and Analysis of Surface Properties. This algorithm can exploit surface connectivity to display internal cavities of proteins. Nicholls (1992) used GRASP to observe that the bRh had numerous "holes" surrounding the retinal moiety. The internal cavities of D_2 was examined (figure 9.4.6). It is clear that a number of cavities exist in the model. Extrapolating this to the real structure presents a number of problems. Firstly, there is currently no 3D x-ray structure for any GPCR. Therefore, a detailed analysis of "holes" in real GPCRs can not be carried out. Secondly, the size and conformity of the "holes" is likely to vary in the real structure. This follows directly from the work of Sylte *et al.* (1993) who performed extensive molecular dynamics (MD) on a model of the 5-HT_{1A} GPCR and ligands. In this work, ligands were observed to move considerable distances inside the receptor and the TM helices positions and tilts changed significantly. Hence, extensive conformational changes in the receptor must impact on the internal cavities in the protein. By dumping structures every picosecond or so - the changing nature of the "holes" could be examined using GRASP. To carry out a detailed analysis of the "holes" in a single model of D_2 is simply not justified.

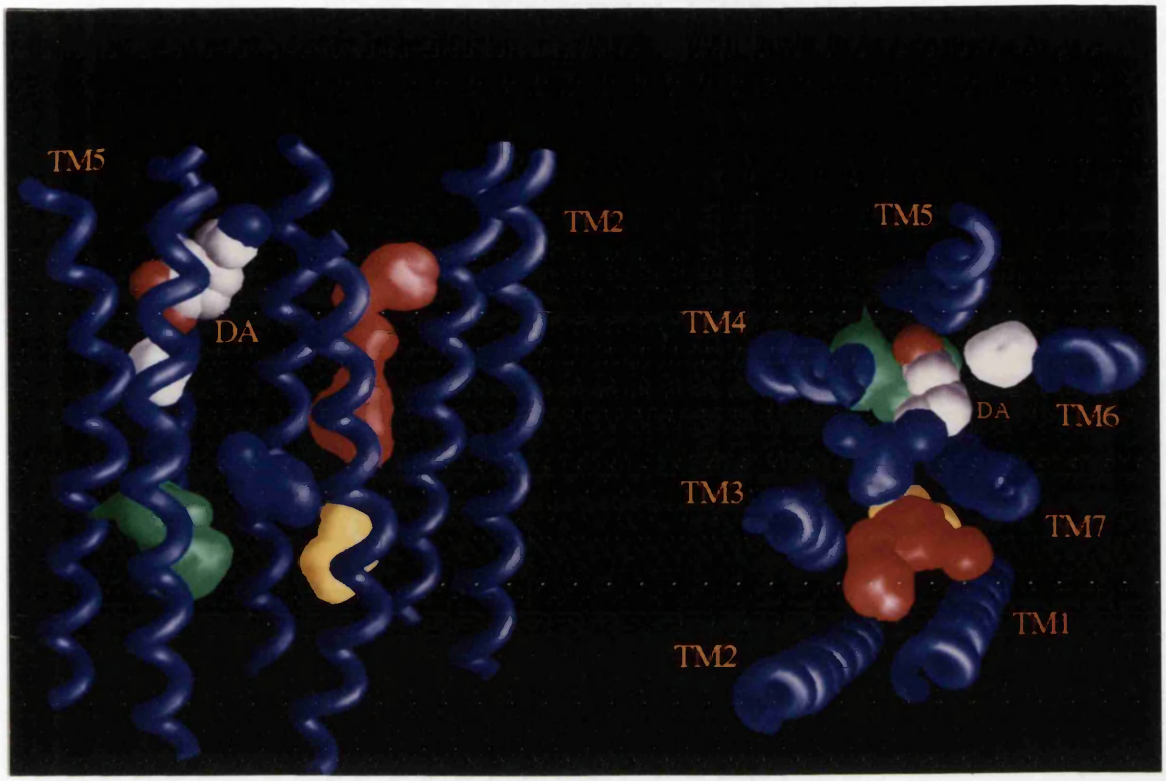


Figure 9.4.6; application of Graphical Representation and Analysis of Surface Properties algorithm (GRASP; Nicholls, 1992) to the 3D model of receptor D_2 incorporating the natural ligand dopamine (DA). The internal cavities of in the model are clearly discernible.

9.4.7 Further Work

With any piece of complicated work the main constraint is time. There is an obvious need to extend the modelling work to include modelling of the extracellular and intracellular domains of the dopamine GPCRs. Also, the modelling of the lipid bilayer around the dopamine GPCR would permit meaningful molecular dynamics to be carried out. However, what is definitely missing from the current work, is a model of the dopamine molecule docked to Asp_{214} which would provide a 3D version of Saunders and Findlay's (1990) MODE 3 - the final target for GPCR agonists.

9.5 Conclusion

All any GPCR modeller can hope to achieve with the absence of a detailed 3D structure of a GPCR is to generate a physically plausible model. That it will not truly reflect the *Real M^{c}Coy* goes without saying. The author is acutely aware that in any modelling exercise the lack of experimental input, from binding studies involving novel ligands to specific mutagenesis experiments, must render the generated models vulnerable to criticism. However, great effort has been taken to scour the available literature to incorporate a number of features in each dopamine model. In particular:

- the likely backbone torsion angles for a helix with a middle Pro (Sankararamakrishnan and Vishveshwara, 1990; Sankararamakrishnan *et al.*, 1991).

- the likely arrangement of transmembrane helices in GPCRs (Baldwin, 1993) based on the experimentally derived low-resolution projection map for rH (Gebhard *et al.*, 1993)

Let's leave the final words to Hibert *et al.* (1993): "This is not a G protein-coupled receptor."

References

Albert, A. (1971). Selective Toxicity and related topics. Fourth edition. London, Chapman and Hall. John Wiley & Sons, New York. Chapter 5; 181-194.

Albert, A. (1979). Selective Toxicity - The physico-chemical basis of therapy. Sixth edition. London, Chapman and Hall. John Wiley & Sons, New York. Chapter 8; 277-288.

Allan, R. (1992). PARIX - The Great Challenge? In Allan. R., Blake, R., Guest, M., Harrison, N., Marshall, S., Rendall, A., Walmsley, P. and Yeung, F. (eds), *Parallel News Issue No. 3*. Daresbury Laboratory, Warrington, UK. pp 2-3.

Allard, W. J., Sigal, I. S. and Dixon, R. A. F. (1987). Sequence Of The Gene Encoding The Human M1 Muscarnic Acetycholine Receptor. *Nucleic Acids Research.* **15(24),** 10604-10604.

Allen, F. H. and Kennard, O. (1983). The Cambridge Database Of Molecular Structures. *Perspectives In Computing.* **3(3),** 28-43.

Andersen, P. H., Gingrich, J. A., Bates, M. D., Dearry, A., Falardeau, P., Senogles, S. E. and Caron, M. G. (1990). Dopamine Receptor Subtypes - Beyond The D1/D2 Classification. *Trends Pharm. Sci.* **11(6),** 231-236.

Anfinsen, C. B. (1959). *The Molecular Basis of Evolution.* John Wiley & Sons, New York.

Anfinsen, C. B., Haber, E., Sela, M. and White, F. H. (1961). The Kinetics of Formation of Native Ribonuclease During Oxidation of the Reduced Polypeptide Chain. *Proc. Natl. Acad., Sci. U.S.A.* **47,** 1309-1314.

Anfinsen, C. B. (1973). Principles that Govern the Folding of Protein Chains. *Science.* **181.** 223-230.

Andersen, J. P., Vilsen, B. and MacLennan, D. H. (1992). Functional Consequences of Alterations to Gly^{310}, Gly^{770} and Gly^{801} Located in the Transmembrane Domains of the Ca^{2+}-ATPase of Sarcoplasminc Reticulum. *J. Biol. Chem.* **267(4),** 2767-2774.

Argos, P., Rao, J. K. M. and Hargrave, P. A. (1982). Structural Prediction Of Membrane Bound Proteins. *European J. Biochem.* **128(2-3),** 565-575.

Argos, P. (1987). A Sensitive Procedure to Compare Amino Acid Sequences. *J. Mol. Biol.,* **193,** 385-396.

Attwood, T. K., Eliopolus, E. E. and Findlay, J. B. C. (1991). Multiple sequence alignment of protein families showing low sequence homology: a methodological approach

using database pattern-matching discriminators for G-protein-linked receptors. *Gene.* **98,** 153-159.

Baker, E N. and Hubbard, R. E. (1984). Hydrogen-Bonding In Globular Proteins. *Prog. Biophys. Mol. Biol.* **44(2),** 97-179.

Baldwin, J. M. (1993). The probable arrangement of the helices in G protein-coupled receptors. *The EMBO Journal.* **12(4),** 1693-1703.

Bangham, J. A. (1988). Data-Sieving Hydrophobicity Plots. *Anal. Biochem.* **174(1),** 142-145.

Barlow, D. J. and Thornton, J. M. (1988). Helix Geometry in Proteins. *J. Mol. Biol.* **201(3),** 601-619.

Bell, L. (1992). Personal communication.

Benkirane, S., Arbilla, S. and Langer, S. Z. (1987). A functional response to D_1 dopamine receptor stimulation in the central nervous system inhibition of the release of [3H]-serotonin from the rat *substantia nigra. Naunyn Solmiedebergs Arch Pharmacol.* **335(5),** 502-507.

Benovic, J. L., Strasser, R. H., Mayor, F., Caron, M. G. and Lefkowitz, R. J. (1986a). β-Adrenergic Receptor Kinase - A Novel cAMP Independent Protein Kinase Which Phosphorylates And Desensitizes The Agonist Occupied Form Of The Receptor. *Clinical Research.* **34(2),** Meeting Abstract.

Benovic, J. L., Strasser, R. H., Caron, M. G. and Lefkowitz, R. J. (1986b). β-Adrenergic Receptor Kinase - Identification Of A Novel Protein Kinase That Phosphorylates The Agonist Occupied Form Of The Receptor. *Proc. Nat. Acad. Sci. U.S.A.* **83(9),** 2797-2801.

Bergen, R. and Carlström, D. (1968). The Structure of the Catecholamines. II. The Crystal Structure of Dopamine Hydrochloride. *Acta Crystallogr.*,**Section B, 24,** 1506-1510.

Beveridge, D. L. and Radna, R. J. (1971). Structural Chemistry Of Cholinergic Neural Transmission Systems. I. A Quantum Theoretical Study Of The Molecular Electronic Structure Of Acetycholine. *J. Amer. Chem. Soc.* **93,** 3759-3764.

Bloom, F. E., Cowey, A., Faull, R. L. M., Grahamsmith, D. G., Gray, J. A., Hokfelt, T., Iversen, L. L., Llinas, R. R., Marsden, C. D., Purpural, D. P., Ratt, M. C., Smith, A. D., Snyder, S. H. and Somogyi, P. (1992). Cost of Brain Disorders. *Nature.* **358(6383),** p348.

Blout, E. R., deLozé, C., Bloom, S. M. and Fasman, G. D. (1960). The Dependence Of The Conformations Of Synthetic Polypeptides On Amino Acid Composition. *J. Amer. Chem. Soc.* **82,** 3787-3789.

Blout, E. R. (1962). In: *Polyamino Acids, Polypeptides and Proteins*. Stahmann, M. A. (Ed.). University of Wisconsin Press, Madison. p275.

Blum, K, Noble, E. P., Sheridan, P. J., Montgomery, A, Ritchie, T., Jagadeeswarran, P., Nogami, H., Briggs, A. H., and Cohn, J. B. (1990). Allelic association of human dopamine D_2 receptor gene in alcoholism. *JAMA*. 263, 2055-2060.

Blundell, T., Barlow, D., Borlakoti, N. and Thornton, J. M. (1983). Solvent induced distortions and the curvature of α-helices. *Nature*. **306,** 281-283.

Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E. and Thornton, J. M. (1987). Knowledge-based prediction of protein structures and the design of novel molecules. *Nature*. **326.** 347-352.

Blundell, T. L. and Johnson, M. S. (1993). Catching the common fold. *Protein Science*. **2,** 877-883.

Boehncke., K., (1990). Molecular dynamics simulations on a systolic ring of transputers. In Wagner, A.S. (ed.). *Transputer Research and Applications 3*. IOS Press, Amsterdam, pp. 83-94.

Bonner, T. I., Young, A. C., Brann, M. R. and Buckley, N. J. (1988). Cloning and Expression of the Human and Rat M5 Muscarnic Acetylcholine-receptor Genes. *Neuron*. **1(5),** 403-410. SWISS-PROT Fetch Number: P08912.

Bordo, D. and Argos, P. (1990). Evolution of Protein Cores - Constraints in Point Mutations as Observed in Globin Tertiary Structures. *J. Mol. Biol.* **211,** 975-988.

Bormann, B. J. and Engelman, D. M. (1992). Intramembrane Helix-Helix Association In Oligomerization and Transmembrane Signaling. *Annu. Rev. Biophys. Biomol. Struct.* **21,** 223-242.

Brandl C. J., Green, N. M., Korczak, B and Maclenan, D. H. (1986). 2 CA-2 and ATPase Genes - Homologies and Mechanistic Implications of Deduced Amino-Acid Sequences. *Cell*. **44(4),** 597-607.

Brucke, T., Wenger, S, Fodreka, I. and Asenbaum, S. (1991). Dopamine Receptor Classification, neuroanatomical distribution and in *vivo* imaging. *Wien Klin Wochenschr*. **103(21),** 639-646.

Burgen, A. S. V., Roberts, G. C. K. and Feeney, J. (1975). Binding of flexible ligands to macromolecules. *Nature*. **253,** 753-755.

Burkhard, R. and Sander, C. (1993). Improved Prediction Of Protein Secondary Structure At Better Than 70% Accuracy. *J. Mol. Biol.* **232,** 584-599.

Burley, S. K. and Petsko, G. A. (1988). Weakly Polar Interactions in Proteins. *Adv. Protein Chem.* **39,** 125-189.

Burley, S. K. and Petsko, G. A. (1989). Electrostatic Interactions In Aromatic Oligopeptides Contribute To Protein Stability. *Trends Biotech.* **7(12),** 354-359.

Bustard, T. M. and Egan, R. S. (1971). The Conformation Of Dopamine Hydrochloride. *Tetrahedron.* **27,** 4457-4469.

Canfield, R. E. and Anfinsen, C. B. (1963). Nonuniform Labeling of Egg White Lysozyme. *Biochemistry.* **2,** 1073.

Cannon, J. G. (1985). *Prog. Drug. Res.* **29,** 303-414.

Carlson, A., Lindqvist, M. L. and Magnusson, T. (1957). 3,4-Dihydroxyphenylalanine and 5-Hydroxytryptophan as Reserpine Antagonists. *Nature,* **180,** 1200.

Carlstrom, D. (1973). *Acta Crystallograp.* Sect. B. **29,** 161.

Carlstrom, D. and Bergin, R. (1967). *Acta Crystallograp.* **23,** 313.

Carver, J. P. and Blout, E. R. (1967). Treatise on Collagen, 1, Academic Press, New York, 441-526.

Charles, D. (1992). The thinking machine's guide to computing. In Richard Fifield (Executive Editor). *New Scientist.* IPC Magazines, London. **1837,** 26-30.

Chothia, C., Lesk, A. M., Dodson, G. G. and Hodgkin, D. C. (1983). Transmission of conformational change in insulin. *Nature.* **302,** 500-505.

Chou, P. Y. and Fasman, G. D. (1974a). Conformational parameters for amino acids in helical, β-sheet and random coil regions calculated from proteins. *Biochemistry.* **13,** 211-222.

Chou, P. Y. and Fasman, G. D. (1974b). Prediction of protein conformation. *Biochemistry.* **115,** 135-175.

Chou, P. Y. and Fasman, G. D. (1977). β-Turns in proteins. *J. Mol. Biol.* **115,** 135-175.

Chou, P. Y. and Fasman, G. D. (1978). Prediction Of The Secondary Structure Of Proteins From Their Amino Acid Sequence. *Adv. Enz.* **47,** 45-147.

Chou, P. Y. and Fasman, G. D. (1979). Prediction of β-Turns. *Biophys. J.* **26,** 367-384.

Chung Fu-Zon, Lentes, K.U., Gocayne, J., Fitzergerald, M, Robinson, D, Kerlavage, A.R., Fraser, C.M. and Venter, J.C. (1987). Cloning and sequence analysis of the human brain beta-adrenergic receptor - Evolutionary relationships to rodent and avain beta-receptors and porcine muscarinic receptors. *FEBS Letters,* **211(2),** 200-206.

Civelli, O., Bunzow, J., Albert, P., Hubert, H. M., Van Tol and Grandy, D. (1992). The Dopamine D_2 Receptor. In: *Molecular Biology of Receptors That Couple To G-Proteins. Applications of Molecular Genetics to Pharmacology.* Editor: Brann, M. R. Birkhauser, Switzerland. ISBN: 3764334657. Chapter 7. Pages 160-169.

Clark, D. and White, F. J. (1987). Review - D1 Dopamine Receptor - The Search For A Function - A Critical Evaluation of the D1/D2 Dopamine Receptor Classification And It's Functional Implications. *Synapse.* **1(4),** 347-388.

Clarke, D. M., Loo, T. W., Inesi, G. and MacLennan, D. H. (1989). Location of high-affinity Ca^{2+} binding sites within the predicted transmembrane domain of the sarcoplasmic-reticulum Ca^{2+}-ATPase. *Nature.* **339,** 476-478.

Cohen, F. E., Richmond, T. J. and Richards, F. M. (1979). Proteins Folding: Evaluations of some Simple Rules for the Assembly of Helices into Tertiary Structures with Myoglobin as an Example. *J. Mol. Biol.,* **132,** 275-293.

Cohen, F. E., Sternberg, M. J. E. and Taylor, W. R. (1980). Analysis And Prediction Of Protein β-Sheet Structures By A Combinatorial Approach. *Nature.* **285,** 378-382.

Cohen, F. E., Abarbanel, R. M., Kuntz, I. D. and Fletterick, R. J. (1983). Secondary Structure Assignment for α/β Proteins by a Combinatorial Approach. *Biochemistry.* **22,** 4894-4904.

Coleman, R. J., Jenner, P., Marsden, C. D. and Stahl, S. M. (1988). Neurol. Neurobiol., **42C,** (*Progr. Catecholamine Res.,* Part C), 19-23.

Cools, A. R. and van Rossan, J. M. (1976). Excitation-Mediating And Inhibition-Mediating Dopamine-Receptors: A New Concept Towards A Better Understanding Of Electrophysciological, Biochemical, Pharmacological, Functional And Clinical Data. *Psychopharmacology.* **45,** 243-254.

Cools, A. R. and van Rossum, J. M. (1980). Multiple Receptors For Brain Dopamine In Behaviour Regulation: Concept Of Dopamine-E And Dopamine-I Receptors. *Life Sci.* **27,** 1237-1253.

Cools, A. R. (1981). The Puzzling Cascade Of Multiple Receptors For Dopamine - An Appraisal of the Current Situtation. *Trends Pharm. Sciences.* **2(7),** 178-183.

Cooper, R.K. and Allan, R.J. (1992). Fortnet (3L) v1.0 - A message passing harness for transputers using 3L Parallel Fortran. *The Transputer Consortium Mailshot.* Rutherford Appleton Laboratory, UK. September Issue, pp 25-36.

Cornette, J. L., Cease, K. B., Margalit, H., Spouge, J. L., Berzofsky, J. A. and DeLisi, C. (1987). Hydrophobiciity Scales And Computational Techniques For Detecting Amphipathic Structures In Proteins. *J. Mol. Biol.* **195(3),** 659-685.

Corpet, F. (1988). Multiple sequence alignment with hierarchal clustering. *Nucleic Acids Res.,* **16,** 10881-10890.

Cotmore, S. F., Furthmay, H., and Marchesi, V. (1977). Immunological evidence for the transmembrane orientation of glycophorin A. Location of ferrin-antibody conjugates in intact cells. *J. Mol. Biol.* **113,** 539.

Cowan, S.W., Schirmer, T., Rummel, G., Steiert, M., Ghosh, R., Pautit, R. A., Jansonius, J. N. and Rosenbusch, J. P. (1992). Crystal-Structures Explain Functional-Properties Of 2 *Escherichia coli* Porins. *Nature,* **358(6389),** 727-733.

Coyle, J. T. (1992). Cost of Brain Disorders. *Nature.* **358,** no. 6383, page184.

Cronet, P., Sander, C. and Vriend, G. (1993). Modelling of transmembrane seven helix bundles. *Prot. Eng.* **6(1),** 59-64.

Crow, T. J. (1994). Prenatal Exposure To Influenza As A Cause Of Schizophrenia - There Are Inconsistencies And Contradictions In The Evidence. *British J. Psychiatry.* **164,** 588-592.

Crow, T. J. and Harrington, C. A. (1994). Etiopathogenesis And Treatment Of Psychosis. *Annual Review of Medicine.* **45,** 219-234.

Dahl, S. G., Edvardsen, O. and Sylte, I. (1991a). Molecular dynamics of dopamine at the D2 receptor. *Proc. Natl. Acad. Sci.USA.* **88,** 8111-8115.

Dahl, S. G., Edvardsen, O. and Sylte, I. (1991b). Molecular modelling of antipsychotic drugs and G protein coupled receptors. *Therapie.* **46,** 453-459.

DaoPin, S., Baase, W. A. and Matthews. B. W. (1990). A mutant T4 lysozyme (Val 131 to Ala 131) designed to increase thermostability by the reduction of strain within an α-helix. *Proteins Struct. Funct. Genet.* **7(2),** 198-204.

Dal-Toso, R., Sommer, B., Ewert, M., Herb, A., Pritchett, D. B., Bach, A., Shivers, B. D. and Seeburg, P. H. (1989). The Dopamine D2 Receptor - 2 Molecular Forms Generated By Alternating Splicing. EMBO. 8, 4025-4034. SWISS PROT FETCH NUMBER: P14416.

Dauber-Osguthorpe, P., Roberts, V. A., Osguthorpe, D. J., Wolff, J., Genest, M. and Hagler, A. T. (1988). Structure and Energetics of Ligand Binding to Proteins: *Escherichia coli* Dihydrofolate Reductase-Trimethoprim, A Drug-Receptor System. *Proteins: Structure, Function, and Genetics.* **4(1),** 31-47.

Davies, D. R. (1964). A Correlation Between Amino-Acid Composition And Protein Structure. *J. Mol. Biol.* **9,** 605-609.

Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978). In Dayhoff, M.O. (ed). *Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, Silver Spring, MD. **5,** suppl. 3, pp. 345-352.

De Lean, A., Stadel, J. M. and Lefkowitz, R. J. (1980). A Ternary Complex Model Explains The Agonist-Specific Binding Properties Of The Adenylate Cyclase Coupled β-Adrenergic Receptor. *J. Biol. Chem.* **255,** 7108-7117.

Dearry, A., Gingrich, J. A., Falardeau, P., Fremeau, R. T., Bates, J. R. and Caron, M. G. (1990). Molecular Cloning and Expression Of The Gene For A Human D1 Dopamine Receptor. Nature. 347, 72-76. SWISS PROT FETCH NUMBER: P21728.

Deber, C. M., Brandl, C. J., Deber, R. B., Hsu, L. C. and Young, X. K. (1986). Amino acid composition of the membrane and aqueous domains of integral membrane proteins. *Arch. Biochem. Biophys.* **251(1),** 68-76.

Deber, M., Glibowicka, M. and Woolley, G. A. (1990). Conformations of Proline Residues in Membrane Environments. *Biopolymers.* **29,** 149-157.

Deber, C. M., Li, Z., Joensson, C., Glibowicka, M. and Xu, G. (1992). Transmembrane Region of Wild-type and Mutant M13 Coat Proteins - Conformational Role of β-branched Residues. *J. Biol. Chem.* **267(8),** 5296-5300.

Deber, C. M., Khan, A. R., Zuomei, L., Joensson, C. and Glibowicka, M. (1993). Val to Ala mutations selectively alter helix-helix packing in the transmembrane segment of phage M13 coat protein. *Proc. Natl. Acad. USA.* **90,** 11648-11652.

Deisenhofer, J., Epp, O., Miki, K., Huber, R., and Michel, H. (1984). X-Ray Structure Analysis Of A Membrane Protein Complex Electron Density Map At 3 Å Resolution And A Model Of The Chromophores Of The Photo-Synthetic Reaction Center From *Rhodopseudomonas viridis. J. Mol. Biol.* **180,** 385-398.

Deisenhofer, J., Epp, O., Miki, K., Huber, R., and Michel, H. (1985). Structure Of The Protein Subunits In The Photosynthetic Reaction Center Of *Rhodopseudomonas viridis* At 3 Å resolution. *Nature.* **318(6047),** 618-624.

De-Keyser, J. (1993). Subtypes And Localization Of Dopamine Receptors In Human Brain. *Neurochemistry International.* **22(2),** 83-93.

Devereux, J., Haeberli, P. and Smithies, O. (1984). A comprehensive set of sequence analysis programs for the Vax. Nucleic Acid Res. **12(1),** 387-395.

Dill, K. A. (1985). Folding and Stability of Globular-Proteins. *Biochem.* **24,** 1501-1509.

Dintzis, H. M. (1961). Assembly of the Peptide Chains of Hemoglobin. *Proc. Nat. Acad. Sci. U.S.A.* **47,** 247.

Dixon, R. A. F., Sigal, I. S., Candelore, M. R., Register, R. B., Scattergood, W., Rands, E. and Strader, C. D. (1987a). Structural features required for ligand binding to the β-adrenergic receptor. *EMBO J.* **6(11),** 3269-3275.

Dixon, R. A. F., Irving, S., Sigal, I. S., Rands, E., Register, R. B., Candelore, M. R., Blake, A. D. and Strader C. D. (1987b). Ligand binding to the β-adrenergic receptor involves its rhodopsin-like core. *Nature,* **326(6108),** 73-7.

Donnelly, D., Johnson, M. S., Blundell, T. L. and Saunders, J. (1989). An Analysis Of The Periodicity Of Conserved Residues In Sequence Alignments of G-Protein Coupled Receptors - Implications For The Three Dimensional Structure. *FEBS Lett.* **251(1),** 109-116.

Donnelly, D., Overington, J. P., Ruffle, S. V., Nugent, J. H. A. and Blundell, T. L. (1993). Modeling α-Helical Transmembrane Domains - The Calculation And Use Of Substitution Tables For Lipid-Facing Residues. *Protein Sci.* **2(1),** 55-70.

Donnelly, D. and Cogdell, R. J. (1993). Predicting the point at which transmembrane helices protude from the bilayer: a model of the antenna complexes from photosynthetic bacteria. *Prot. Eng.* **8(6),** 629-635.

Donnelly, D., Findlay, J. B. C. and Blundell, T. L. (1994). The Evolution And Structure Of Aminergic G Protein Coupled Receptors. *Receptors & Channels.* **2(1),** 61-78.

Doolittle, R. F. (1981). Similar amino acids: chance or common ancestry? *Science.* **214,** 149-159.

Dougherty, D. A. and Stauffer, D. A. (1990). Acetylcholine Binding by a Synthetic Receptor: Implications for Biological Recognition. *Science.* **250,** 1558-1560.

Dunn, R. J., McCoy, J., Simsek, M., Majumdar, A., Chang, S. H., Rajbhandary, U. L. and Khorana, H. G. (1981). The Bacteriorhodopsin Gene. *Proc. Natl. Sci. U.S.A.* **78,** 6744-6748.

Eisenberg, D., Weiss, R. M. and Terwilliger, T. C. (1984). The Hydrophobic Moment Detects Periodicity In Protein Hydrophobicity. *Proc. Natl. Acad. Sci. U.S.A.* **81,** 140-144.

Emorine, L. J., Maullo, S., Briendsutren, M. M., Patey, G. and Tate, K. (1989). Molecular Characterization Of The Human Beta-3-Adrenergic Receptor. *Science.* **245(4922),** 1118-1121.

Engel, M., Williams, R. W. and Erickson, B. W. (1991). Designed Coiled-Coil Proteins Synthesis and Spectroscopy of 2 78-residue α-helical Dimers. *Biochemistry.* **30(13),** 3161-3169.

Engelman, D. M., Henderson, R., McLachlan, A. D. and Wallace, B. A. (1980). Path of the polypeptide in bacteriorhodopsin. *Proc. Natl. Acad. Sci. USA.* **77(4),** 2023-2027.

Engelman, D.M. Steitz, T.A. and Goldman, A. (1986). Identifying Nonpolar Transbilayer Helices In Amino Acid Sequences Of Membrane Proteins. In Engelman, D.M., Cantor, C.R. and Pollard, T.D. (eds). *Ann. Rev. Biophys. Chem.,* **15,** 321-353.

Evans, S. V. (1993). SECTOR: Hardware - Lighted Three Dimensional Solid Model Representation Of Macromolecules. Version 4.13.0. Manual.

Eylar, E. H. (1965) On the biological role of glycoproteins. *J. Theor. Biol.***10,** p89.

Fasman, G. D. (1989). The Development of the Prediction of Protein Structure. In: *Prediction of Protein Structure and the Principles of Protein Conformation.* Editor: Fasman, G. D. Plenum Press, New York. ISBN: 0-306-43131-9. Chapter 6, 193-316.

Fasman, G. D. and Gilbert, W. A. (1990). The prediction of transmembrane protein sequences and their conformation: an evaluation. *Trends Biochem. Sci.* **15(3),** 89-92.

Feng, D. and Doolittle, R. (1987). Progressive Multiple Sequence Alignment. *J. Mol. Evol.* **25,** 351-360.

Findlay, J. B. C. and Pappin, D. J. C. (1986). *Biochem.J.* **238,** 625-642.

Findlay, J. B. C. and Pappin, D. J. C. (1988). *Prog. Retinal Res,* **7,** 63-87.

Findlay, J. B. C. and Geisow, M. J. - editors. (1989). Protein sequencing a practical approach. IRL PRESS, Oxford. ISBN: 0-19-963012-7.

Findlay, J. and Eliopoulos, E. (1990). Three-dimensional modelling of G protein-linked receptors. *Trends Pharm. Sci.* **11,** 492-499.

Findlay, J.B.C., Eliopoulos, E.E. and Attwood, T.K. (1990). The Structure of G-Protein-Linked Receptors. In Milligan, G., Wakelam, M.J.O. and Kay, J. (eds). G-Proteins And Signal Transduction. *Biochem. Soc. Symp.*, **56,** 1-8.

Findlay, J. (1991). Across the biological membrane. *CHEMISTRY IN BRITAIN.* **27(8),** 724-727.

Finer-Moore, J. and Stroud, R. M. (1984). Amphipathic Analysis And Possible Formation Of The Ion Channel In The Acetylcholine Receptor. *Proc. Nat. Acad. Sci. U.S.A.* **81(1),** 155-159.

Filtz, T. M., Artymyshyn, R. P., Guan, W. and Molinoff, P. B. (1993). Paradoxical Regulation Of Dopamine Receptors In Transfected 293 Cells. *Mol. Pharm.* **44(2),** 371-379.

Fong, T. M., Cascieri, M. A., Yu, H., Bensal, A., Swain, C. and Strader, C. D. (1993). Amino-aromatic interaction between histidine 197 of the neurokinin-1 receptor and CP 96345. *Nature.* **362,** 350-353.

Fraser, C. M., Arakawa, S., McCombie, W. R. and Venter, J. C. (1989). Cloning, Sequence Analysis And Permanent Expression Of A Human Alpha-2-Adrenergic Receptor In Chinese Hamster Ovary Cells - Evidence For Independent Pathways Of Receptor Coupling To Adenylate Cyclase Attenuation And Activation. *J. Biol. Chem.* **264(20),** 11754-11761.

Frielle, T., Collins, S., Daniel, K. W., Caron, M. G., Lefkowitz, R. J. and Kobilka, B. K. (1987). Cloning Of The cDNA For The Human Beta-1-Adrenergic Receptor. *Proc. Natl. Acad. Sci. USA.* **84(22),** 7920-7924.

Frielle, T., Daniel, K. W., Caron, M. G., Lefkowitz, R. J. (1988). Structural Basis Of β-Adrenergic Receptor Specificity Studied With Chimeric β_1/β_2-Adrenergic Receptors. *Proc. Natl. Acad. Sci. USA.* **85,** 9494-9498.

Friemann, A. and Schmitz, S. (1992). A new approach for displaying identities and differences among aligned amino acid sequences. *Comput. Applic. Biosci.,* **8,** 261-265.

Fujiwara, Y, Sura, I. and Tomila, H. (1991). The Molecular Biology of Dopamine Receptors. *Yakubulsu-Seishin-Kodo.* **11(8),** 187-196.

Garavito, R. M. (1990). Crystallization of Membrane-Proteins. *Biophysical Journal.* **57(2).** Meeting abstract.

Garnier, J., Osguthorpe, D. J. and Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120,** 97-102.

Garnier, J. and Levin, J. M. (1991). The Protein Structure Code - What Is Its Present Status. *CABIOS.* **7(2),** 133-142.

Garnier, J. and Robson, B. (1989). The GOR Method for Predicting Secondary Structures in Proteins. In: *Prediction of Protein Structure and the Principles of Protein Conformation.* Editor: Fasman, G. D. Plenum Press, New York. ISBN: 0-306-43131-9. Chapter 10, 417-464.

Gebhard F. X., Schertler, Claudio Villa and Henderson, R. (1993). Projection structure of rhodopsin. Nature. 362, 770-772.

Geisow, M. J. and Roberts, R. D. B. (1980). Amino acid preferences for secondary structure vary with protein class. *Int. J. Biol. Macromol.* **2,** 387-389.

Gerhard, U., Searle, M. S. and Williams, D. H. (1993). The Free Energy Change Of Restricting A Bond Rotation In The Binding Of Peptide Analogs To Vancomycin Group Antibiotics. *Bioorganic & Med. Chem Lett* **3(5),** 803-808.

Gerstein, M.and Chothia, C. (1991). Analysis of Protein Loop Closure - Two Types of Hinges Produce One Motion in Lactate Dehydrogenase. *J. Mol. Biol.* **220,** 133-149.

Gibrat, J. F. (1986). Modelisation By Computer Of The 3D Structure Of Proteins. Ph.D. Thesis. University of Paris VI Paris.

Gibrat, Jean-Francois, Robson, B. and Garnier, J. (1991). Influence of the Local Amino Acid upon the Zones of the Torsional Angles φ and ψ Adopted by Residues in Proteins. *Biochemistry.* **30,** 1578-1586.

Giesecke, J. (1980). Refinement of the Structure of Dopamine Hydrochloride. *Acta Crystallogr.* **Section B, 36,** 178-181.

Goldberger, R. F., Epstein, C. B., Anfinsen, C. B. (1963). Acceleration Of Reaction Of Reduced Bovine Pancreatic Ribonucleases By A Microsomal System From Rat Liver. *J. Biol Chem.* **238,** 628-635.

Goldsack, D. E. (1969). Relation of Amino Acid Composition And The Moffitt Parameters For The Secondary Structure Of Proteins. *Biopolymers.* **7,** 299-313.

Goodfellow J.M. (1990). Computer Simulation in Molecular Biology. *Chemistry In Britain.* **26,** 1066-1068.

Goodfellow J.M., Jones, D.M., Laskowski, R.A., Moss, D.S., Saqi, M., Thanki, N. and Westlake, R. (1990). Use of Parallel Processing In The Study of Protein Ligand Binding. *J. Comput. Chem.,* **11,** 314-325.

Gonzalez, J.C. and Lopez, J. (1991). A Probabilistic Expert System for Diagnosis of Skin Diseases. In Durrani, T.S., Sandham, W.A., Soraghan, J.J. and Forbes S.M. (eds.), Applications of Transputers 3 - Volume 2. *Proceedings of the Third International Conference on Applications of Transputers.* IOS Press, Oxford, UK. 433-438.

Green, B. R. (1990). Structure prediction methods for membrane proteins: comparisons with the x-ray structure of the *R. Viridis* photosynthetic reaction centre. *Current Reseach in Protein Chemistry: Techniques, Structure and Function.* Editor: Villafranca, J. J. Academic Press, San Diego. ISBN: 0-12-721956-0. Chapter 36, pp 395-404.

Greer, J. (1981). Comparative Model-Building of the Mammalian Serine Proteases. *J. Mol. Biol.* **153,** 1027-1042.

Greer, J. (1990a). Comparitive Modeling Of Proteins In The Design Of Novel Renin Inhibitors. *Biophysical Journal.* **57(2),** meeting abstract.

Greer, J. (1990b). Comparitive Modeling Methods - Applications To The Family Of The Mammalian Serine Proteinases. *Proteins.* **7(4),** 317-334.

Greer, J. (1991). Comparitive Modelling Of Homologous Proteins. *Methods Enzymol.* **202,** 239-252.

Gregoret, L. M. and Cohen, F. E. (1990). Novel method for the rapid evaluation of packing in protein structures. *J. Mol. Biol.* **211,** 959-974.

Gribskov, M. and Devereux, J. (1991). Sequence Analysis Primer. Stockton Press, New York. ISBN 1-56159-007-X. Chapters 3 and 4.

Grol, C. J., Jansen, L. J. and Rollema, H. (1985). Resolution Of 5,6-Dihydroxy-2-(N,N-Di-Normal-Propylamino)tetralin In Relation To The Structural Requirements For Centrally Acting Dopamine Agonists. *J. Med. Chem.* **28(5),** 679-683.

Gunnar von Heijne. (1991a). Proline Kinks In Transmembrane α-Helices. *J. Mol. Biol.* **218,** 499-503.

Gunnar, von Heijne. (1991b). Computer analysis of DNA and protein sequences. *Eur. J. Biochem.,* **199,** 253-256.

Harris, G. C. and Aston-Jones, G. (1994). Involvement Of D_2 Dopamine Receptors In The Nucleus Accumbens In The Opiate Withdrawal Syndrome. *Nature.* **371,** 155-157.

Hayes, G., Biden, T. J., Selbie, L. A. and Shine, J. Structural subtypes of the dopamine D2 receptor are functionally distinct: expression of the cloned D2A and D2B subtypes in a heterologous cell line. (1992). *Mol-Endocrinol.* **6(6),** 920-926.

Hecht, M. H., Richardson, J. S., Richardson, D. C. and Ogden, R. C. (1990). *De Novo* Design, Expression, and Characterization of Felix: A Four-Helix Bundle Protein of Native-Like Sequence. *Nature.* **249,** 884-891.

Henderson, R. and Unwin, P. N. T. (1975). Three-Dimensional Model of Purple Membrane Obtained By Electron Microscopy. *Nature.* **257,** 28-32.

Henderson, R., Baldwin, J., Ceska, T. H., Zemlin, F., Beckmann, E. and Downing, K. (1990). Model of the structure of bacteriorhodopsin based on high resolution electron cryo-microscopy. *J Mol Biol.* **213,** 899-929.

Hibert, M. F., Trump-Kallmeyer, S., Hoflack, J. and Bruinvels, A. (1993). This is not a G protein-coupled receptor. *Trends Pharm. Sci.* **14,** 7-12.

Hoflack, J., Trumpp-Kallmeyer, S. and Hibert, M. (1994). Re-evaluation of bacteriorhodopsin as a model for G protein-coupled receptors. *Trends Pharm. Sci.* **15,** 7-9.

Horn, A. S., Korf, J. and Westerink, R. H. C. (eds). 'The Neurobiology of Dopamine', Academic Press, London. (1979).

Horn, A. S. (1990). Dopamine Receptors. Dopamine Receptors. In: Comparitive Medicinal Chemistry - *The Rational Design, Mechanistic Study & Therapeutic Application Of Chemical Compounds.* Joint Executive Editors: Sammes, P. G. and Taylor, J. B. Vol 3, Chapter 12.3. Pages 229-290.

Hubert, H. M., Tol, V., Wu, C. M., Guan, H., Ohara, K., Bunzow, J. R., Civelli, O., Kennedy, Seeman, P., Niznik, H. B. and Jovanovic, V. (1992). Multiple Dopamine D_4 Receptor Variants In The Human Population. *Nature.* **358,** 149-152.

Humbler, C. and Mizadegan, T. (1992). Three-dimensional Models of G-Protein Coupled Receptors. *Annual Reports In Medicinal Chemistry.* **27,** Chapter 30, 291-300.

IBC (1989) Symposium on Schizophrenia. 5-HT_3 antagonists and ligands for dopamine D_1 and autoreceptors offer new leads for antipsychotic drugs. Meeting Report. *Trends Pharm. Sciences.* (1990). **11,** 49-51.

Ijzerman A. P., Vanderwenden E. M., Vangalen P.J.M., Jacobson K.A. (1994). Molecular Modeling Of Adenosine Receptors - The Ligand-Binding Site On The Rat Adenosine A(2a) Receptor. *European Journal Of Pharmacology-Molecular Pharmacology Section.* **268(1),** 95-104.

Ikezu, T., Okamoto, T., Ogata, E. and Nishimoto, I. (1992). Amino Acids 356-372 Constitute A G_i-activator Sequence Of The α_2-Adrenergic Receptor And Have A Phe Substitute In The G Protein-activator Sequence Motif. *FEBS.* **311(1),** 29-32.

Inmos (1988) *Occam 2 Reference Manual.* Prentice Hall, London.

Jähnig, F. (1989). Structure Prediction for Membrane Proteins. In: *Prediction of Protein Structure and the Principles of Protein Conformation.* Editor: Fasman, G. D. Plenum Press, New York. ISBN: 0-306-43131-9. Chapter 18, 707-717.

Jähnig, F. (1990). Structure Predictions Of Membrane-Proteins Are Not That Bad. *Trends Biochem. Sci.* **15(3),** 93-95.

Jameson, B. A. (1989). Modelling in peptide design. *Nature.* **341,** 465-466.

Johnson, M. S., Overington, J. P. and Šali, A. (1990). Knowledge-Based Protein Modelling: Human Plasma Kallikrein and Human Neutrophil Defensin. In: *Current Research in Protein Chemistry: Techniques, Structure, and Function.* Editor: Villafranca, J.J. Academic Press, London - Published under the Auspices of the Protein Society. Section VI - Structure Prediction Workshop. Page 557-565. ISBN 0-12-721956-0.

Johnson, M. S., Srinivasan, N., Sowdhamini, R. and Blundell, T. L. (1994). Knowledge-Based Protein Modelling. *Critical Reviews in Biochemistry and Molecular Biology.* **29(1),** 1-68.

Johnston, J. M., Wood, D. F., Read, S. and Johnston, D. G. (1993). Dopamine regulates D_2 receptor gene expression in normal but not in tumorous rat pituitary cells. *Mol. Cell. Endocrinol.* **92(1),** 63-68.

Jones, D.M. and Goodfellow, J.M. (1990). In Pritchard D. Scott C. (eds.), *Transputer Applications 90.* IOS Press, Amsterdam.

Jones, R. (1992). Sequence pattern matching on a massively parallel computer. *CABIOS.* **8,** 377-383.

Jose, F. A., Raymond, J. R., Bates, M. D., Aperia, A., Felder, R. A. and Carey, R. M. (1992). The Renal Dopamine Receptors. *J. Am. Soc. Nephrol.* **2(8),** 1265-1278.

Junk, K. (1990). Automatic Beginners Conversion For Parallelism? *Parallelogram International - April Issue.* Parallel Publishing Ltd., London. pp 18-19.

Junk, K. (1991). Software Tools For Transputers - Little Work, With Plantage. *Parallelogram International. - June/July Issue.* Parallel Publishing Ltd., London. pp 16-18.

Kaiser, C. and Jain, T. (1985). Dopamine Receptors - Functions and emerging concepts. *Med. Res. Rev.* **5(2),** 145-229.

Kawai, H., Kikuchi, T. and Okamoto, Y. (1989). A prediction of tertiary structures of peptide by the Monte Carlo simulated annealing method. *Protein Eng.* **3(2),** 85-94.

Klein, C. L. (1991). Experimental Electron Density Distribution of Dopamine Hydrochloride. *Struct. Chem.* **2(5),** 507-514.

Klotz, I. M., Klippenstein, G. L. and Hendrickson, W. A. (1976). Hemerythrin: Alternative oxygen carrier. *Science.* **192,** 335-344.

Knowles, J. R. (1987). Tinkering With Enzymes - What Are We Learning. *Science.* **236(4806),** 1252-1258.

Kobilka, B. K., Frielle, T., Collins, S., Yang-Feng, T., Kobilka, T. S., Francke, U., Lefkowitz, R. J. and Caron, M. G. (1987). An Intronless Gene Encoding A Potential Member Of The Family Of Receptors Coupled To Guanine - Nucleotide Regulatory Proteins. *Nature.* **329(6134),** 75-79.

Komiya, H., Yeates, T. O., Rees, D. C., Allen, J. P. and Feher, G. (1988). Structure Of The Reaction Center From *Rhodobacter sphaeroides* R-26 And 2.4.1-Symmetry-Relations And Sequence Comparisons Between Different Species. *Proc. Natl. Acad. Sci. USA.* **85(23),** 9012-9016.

Kulkarni, R.V., Date, S.T., Kulkarni, B., Kulkarni, U. and Kolaskar, A.S. (1991). PRAS - Parallel Alignment of Sequences. In Durrani, T.S., Sandham, W.A., Soraghan, J.J. and Forbes S.M. (eds.). Applications of Transputers 3 - Volume 2. *Proceedings of the Third International Conference on Applications of Transputers.* IOS Press, Oxford, UK. pp 439.

Kunapuli, P., Gurevich, V. V. and Benovic, J. L. (1994). Phospholipid Stimulated Autophosphorylation Activates The G Protein Coupled Receptor Kinase GRK5. *J. Biol. Chem.* **269(14),** 10209-10212.

Kyte, J. and Doolittle, R. F. (1982). A Simple Method for Displaying the hydropathic Character of a Protein. *J Mol Biol.* **157,** 105-132.

Landolt-Marticorena, C., and Williams, K. A., Deber, C. M. and Reinhard, A. F. (1993). Non-Random Distribution Of Amino-Acids In The Transmembrane Segments Of Human Type I Single Span Membrane Proteins. *J. Mol. Biol.* **229(3),** 602-608.

Lapthorn, A. J., Harris, D. C., Littlejoin, A., Lustbader, J. W., Canfield, R. E., Machin, K. J., Morgan, F. J. and Isaacs, N. W. (1994). Crystal Structure Of Human Chorionic Gonadotropin. *Nature,* **369 (6480),** 455-461.

Laskowski, R. A., Macarthur, M. W., Moss, D. S., Thornton, J. M. (1993). PROCHECK - A Program To Check The Steriochemical Quality Of Protein Structures. *J. App. Crystallography.* **26,** 283-291.

Lemmon, M. A. and Engelman, D. M. (1991). Helix-helix interactions inside lipid bilayers. *Current Opinion in Structural Biology.* **2,** 511-518.

Lemmon, M. A., Flanagan, J. M., Hunt, J. F., Adair, B. D., Borman, B. J., Dempsey, C. E and Engelman, D. M. (1992a). Glycophorin A Dimerization Is Driven by Specific Interactions between Transmembrane α-Helices. *J. Biol. Chem.* **267(11),** 7683-7689.

Lemmon, M. A., Flanagan, J. M., Treutlein, H. R., Zhang, J. and Engelman, D. M. (1992b). Sequence Specificity in the Dimerization of Transmembrane α-Helices. *Biochemistry.* **31,** 12719-12725.

Lemmon, M. A., Treutlein, H. R., Adams, P. D., Brunger, A. T. and Engelman, D. M. (1994). A Dimerization Motif For Transmembrane Alpha-Helices. *Nature Structural Biology,* **1(3),** 157-163.

Lesk, A. M. and Chothia, C. (1984). Mechanisms of Domain Closure in Proteins. *J. Mol. Biol.* **174,** 175-191.

Levinthal, C. in *Mossbauer Spectorscopy in Biological Systems* (eds. Debrunner, P., Tsibris, J. C. M. and Münck, E.) 22-44 (Univ. Illinois Press, Urbana, 1969).

Leysen, J. E., Janssen, P. M. F., Megens, A. A. H. P. and Schoffe, A. (1994). Risperidone - A Novel Antipsychotic With Balanced Serotonin-Dopamine Antagonism, Receptor Occupancy Profile And Pharmacological Activity. *J. Clinical Biochem.* **55,** 5-12.

Levitt, M. and Perutz, M. F. (1988). Aromatic Rings Act as Hydrogen Bond Acceptors. *J Mol Biol.* **201,** 751-754.

Li, Shun-Cheng and Deber, C. M. (1992a). Influence of glycine residues on peptide conformation in membrane environments. *Int. J. Peptide Protein Res.* **40,** 243-248.

Li, Shun-Cheng and Deber, C. M. (1992b). Glycine and β-branched residues support and modulate peptide helicity in membrane environments. *FEBS.* **311(3),** 217-220.

Li, Shun-Cheng and Deber, C. M. (1993). Peptide Environment Specifies - Helicity Of Hydrophobic Segments Compared In Aqueous, Organic, And Membrane Environments. *J. Biol. Chem.* **268(31),** 22975-22978.

Libert, F., Parmentier, M., Lefort, A., Dinsarf, C., Vansande, J., Maenhaut, C., Simons, M. J., Dumont, J. E. and Vassart, G. (1989). Selective Amplification And Cloning Of 4 New Members Of The G-Protein Coupled Receptor Family. *Science.* **244 (4904),** 569-572.

Lim, V. I., (1974a). Structural principles of the globular organization of protein chains: A stereochemical theory of globular protein secondary structure. *J. Mol. Biol.* **88,** 857-872.

Lim, V. I., (1974b). Algorithmsfo rprediction of α-helices and β-structural regions in globular proteins. *J. Mol. Biol.* **88,** 873-894.

Lipman, D. J. and Pearson, W. R. (1985). Rapid And Sensitive Protein Similarity Searches. *Science.* **227,** 1435-1441.

Lomasney, J. W., Lorenz, W., Allen, L. F., King, K., Regan, J. W. and Yangfeng, T. L. (1990). Expansion Of The Alpha-2-Adrenergic Receptor Family - Cloning And Characterisation Of A Human Alpha-2-Adrenergic Receptor Subtype, The Gene For Which Is Located On Chromosome-2. *Proc. Natl. Acad., Sci. U.S.A.* **87(13),** 5094-5098.

Lomax, A.J., Ross, P.G.B. and Undrill, P.E. (1991). Parallel Processing Techniques for the Interactive Display of Volumetric Data Sets. In Durrani, T.S., Sandham, W.A., Soraghan, J.J. and Forbes S.M. (eds.). Applications of Transputers 3 - Volume 1. *Proceedings of the Third International Conference on Applications of Transputers.* IOS Press, Oxford, UK. pp 136-141.

Mackay, D. H. J., Cross, A. J. and Hagler, A. T. (1989). The Role of Energy Minimisation in Simulation Strategies of Biomolecular Systems. In: *Prediction of Protein Structure and the Principles of Protein Conformation.* Editor: Fasman, G. D. Plenum Press, New York. ISBN: 0-306-43131-9. Chapter 7, 317-358.

MacKenzie, R. G., Steffey, M. E., Manelli, A. M., Polleck, N. J. and Frail, D. E. (1993). A D1/D2 Chimeric Dopamine Receptor Mediates A D1 Response To A D2 - Selective Agonist. *Febs. Letts.* **323(1-2),** 59-62.

Maddox, J. (1994). Does folding determine protein configuration? *Nature.* **370,** 13.

Malmberg, A., Jackson, D. M., Eriksson, A. and Mohell, N. (1993). Unique Binding Characteristics Of Antipsychotic Agents Interacting With Human Dopamine D(2A), D(2B) And D(2) Receptors. *Mol. Pharm.* **43(5),** 749-754.

MaloneyHuss, K. and Lybrand, T. P. (1992). Three-dimensional Structure for the β_2 Adrenergic Receptor Protein Based on Computer Modelling Studies. *J Mol Biol.* **225,** 859-871.

Marshall, R. D. (1974). The nature and metablolism of the carbohydrate-peptide linkage of glycoproteins. *Biochem. Soc. Symp.* **40,** p17.

Martonosi, A. N. (1985). *The Enzymes of Biological Membranes.* 2nd ed. Plenum Press, New York.

McGillivary, R. T. A., Mendez, E., Shewale, J. G., Sinha, S. K., Linebackzins, J. and Brew, K. (1983). The Primary Structure Of Human-Serum Transferrin - The Structures Of 7 Cyanogen-Bromide Fragments And The Assembly Of The Complete Structure. *J. Biol. Chem.* **258(6),** 3543-3553.

McGregor, M. J., Islam, S. A. and Sternberg, M. J. E. (1987). Analysis of the Relationship Between Side-Chain Conformation and Secondary Structure in Globular Proteins. *J Mol Biol.* **198,** 295-310.

Milner-White, E. J., Bell, L. H. and Maccallum, P. H. (1992). Pyrrolidine Ring Puckering In *Cis* and *Trans*-Proline Residues In Proteins And Polypeptides - Different Puckers Are Favoured In Certain Situations. *J. Mol. Biol.* **228(3),** 725-734.

Montmayeur, J. P., Guiramand J. and Borrelli, E. (1993). Preferential coupling between dopamine D_2 receptors and G-protein. *Mol. Endocrinology.* **7(2),** 161-170.

Mostad, A. and Romming, C. (1974). *Acta Chem. Scand.* Ser B. **28,** 1161.

Nathans, J. and Hogness, D. S. (1984). Isolation And Nucleotide-Sequence Of The Gene Encoding Human Rhodopsin. *Proc. Natl. Acad., Sci. U.S.A.* **81(15),** 4851-4855.

Neal, M. J. (1989). Medical Pharmacology At A Glance. Blackwell Scientific Publications, London. Chapter 2. Pages 10-11. ISBN: 0-632-01345-1.

Needleman, S.B. and Wunsch, C.D. (1970). A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.,* **48,** 443-453.

Neumeyer, J. L., Reischig, D., Anana, G. W., Campbell, A., Baldessarine, R. J., Kula, N. S. and Watling, K. J. (1983). Aporphines .48. Enantioselectivity of (R)-(-)- and (S)-(&)-N-Normal-Propylnorapomorphine On Dopamine Receptors. *J Med Chem.* **26(4),** 516-521.

Nicholls, A. (1992). Manual for: GRASP - Graphical Representation and Analysis of Surface Properties. Version 1.1. Introduction. Page 3-4.

Nicholls, A., Bharadwaj, R. and Honig, B. (1993). GRASP - Graphical Representation And Analysis Of Surface Properties. *Biophys. J.* **64(2),** page 166, Meeting Abstract.

Nicolson, Garth, L. and Singer, S. J. (1974). The Distribution And Asymmetry Of Mammalian Cell Surface Saccharides Utilizing Ferritin-Conjugated Plant Agglutinins As Specific Saccharide Stains. *J. Cell Biol.* **60(1),** 236-248.

Nishimoto, I. and Okamoto, T. (1992). Detection Of G Protein-activator Regions In M_4 Subtype Muscarnic, Cholinergic, And α_2-Adrenergic Receptors Based Upon Characteristics In Primary Structure. *J. Biol. Chem.* **267(12),** 8342- 8346.

Niznik, H. B. (1987). Dopamine Receptor - Molecular Structure And Function. *Mol. Cell. Endocrinology.* **54(1),** 1-22.

Niznik, H. B. and Van Tol, H. H. (1992). Dopamine receptor genes: new tools for molecular psychiatry. *J. Psychiatry-Neurosci.* **17(4),** 158-180.

Niznik, H. B., Sunahara, R. K., Hubert, H. M., Van Tol H. H., Seeman, P., Weiner, D. M., Storeman, T. M., Brann, M. R. and O'Dowd, B. F. (1992). The Dopamine D1 Receptors. In: *Molecular Biology of Receptors That Couple To G-Proteins. Applications of Molecular Genetics to Pharmacology.* Editor: Brann, M. R. Birkhauser, Switzerland. ISBN: 3764334657. Chapter 6. Pages 142-169.

Oprian, D. D. (1992). The Ligand-Binding Domain Of Rhodopsin and Other G Protein-Linked Receptors. *Journal Of Bioenergetics And Biomembranes.* **24(2),** 211-217.

Pain, R. H. and Robson, B. (1970). Analysis Of The Code Relating Sequences To Secondary Structure. *Nature.* **227,** 62-63.

Parallel Fortran User Guide. (1991). 3L Ltd., Livingston, Scotland.

Parallelogram International. - June/July Issue. (1991). Parallel Publishing Ltd., London. pp 3-4.

Park, M. K., Yoo, H. S., Kang, Y. K. and Lee, N. S. (1992). Conformational Analysis Of Catecholamines - Raman, High-Resolution NMR, And Conformationall Energy Calculation Study. *Bulletin Of The Korean Chem. Soc.* **13(3),** 230-235.

Paul, C. and Rosenbusch, J. P. (1985). Folding of Porin and Bacteriorhodopsin. *EMBO J.* **4(6),** 1593-1597.

Pauling, L. and Corey, R. B. (1951). Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheats. *Proc. Nat. Acad. Sci. U.S.A.* **37,** 729-740.

Pauling, L., Corey, R. B. and Branson, H. R. (1951). The Structure Of Proteins: Two Hydrogen-Bonded Helical Configurations Of The Polypeptide Chain. *Proc. Nat. Acad. Sci. U.S.A.* **37,** 205-211.

Peralta, E. G., Ashkenazi, A., Winslow, J. W., Smith, D. H., Ramachandran, J. and Capon, D. J. (1987). Distinct Primary Structures, Ligand-Binding Properties and Tissue-Specific Expression of Four Human Muscarnic Acetylcholine-Receptors. *EMBO Journal.* **6(13),** 3923-3929. SWISS-PROT Fetch Number: P20309.

Perutz, M. F., Fermi, G., Abraham, D. J., Poyart, C. and Bursaux, E. (1986). Haemoglobin As A Receptor Of Drugs And Peptides - X-ray Studies Of The Steriochemistry Of Binding. *J Amer Chem Soc.* **108,** 5, 1064-1078.

Pettersson, I., Gundertofte, K., Palm, J. and Liljefors, T. (1992). A Study Of The Contributions Of The 1-Phenyl Substituent To The Molecular Electrostatic Potentials Of Some Benzazepines In Relation To Selective Dopamine D-1 Receptor Activity. *J. Med. Chem.* **35,** 502-507.

Piela, L., Némethy, G. and Harold, A. (1987). Proline-Induced Contraints in α-Helices. *Biopolymers.* **26,** 1587-1600.

Polinsky, A., Goodman, M., Williams, K. A. and Deber, C. (1992). Minimum Energy Conformations of Proline-Containing Helices. *Biopolymers.* **32,** 399-406.

Ponder, J. W. and Richards, F. M. (1987). Tertiary Templates for Proteins - Use of Packing Criteria in the Enumeration of Allowed Sequences for Different Structural Classes. *J Mol Biol.* **193,** 775-791.

Popot, J. L. and Engelman, D. M. (1990). Membrane-Protein Folding And Oligomerization - The 2 Stage Model. *Biochemistry.* **29(17),** 4031-4037.

Powell, M. J. and Ogden, J. E. (1990). Nucleotide-Sequence Of Human Lactoferrin cDNA. *Nuc. Acids Res.* **18(13),** 4013-4013.

Prevelige, Jr., P. and Fasman, G. D. (1989). Chou-Fasman Prediction of the Secondary Structure of Proteins - The Chou-Fasman-Prevelige Algorithm. In: *Prediction of Protein Structure and the Principles of Protein Conformation.* Editor: Fasman, G. D. Plenum Press, New York. ISBN: 0-306-43131-9. Chapter 9, 391-416.

Probst, W. C., Snyder, L. A., Schuster, D. I., Brostus, J. and Sealfon, S. C. (1992). Sequence Alignment Of The G Protein Coupled Receptor Superfamily. *DNA Cell Biol.* **11(1),** 1-20.

Ptitsyn, O. B. and Finkelstein, A. V. (1983). Theory of Protein Secondary Structure and Algorithm of Its Prediction. *Biopolymers.* **22,** 15-25.

Raine, A.R.C., Fincham, D. and Smith, W. (1989). Systolic loop methods for molecular dynamics simulations using a ring of transputers. *Comput. Phys. Commun.,* **55,** 13-30.

Rea, W.N. (1991). Performance of Task-Farming With Transputers. In Durrani, T.S., Sandham, W.A., Soraghan, J.J. and Forbes S.M. (eds.), Applications of Transputers 3 - Volume 2. *Proceedings of the Third International Conference on Applications of Transptuers.* IOS Press, Oxford, UK. pp 792-797.

Reddaway, S. (1992). Letters. In Richard Fifield (Executive Editor). *New Scientist.* IPC Magazines, London. **1840**. page 52.

Rees, D. C., DeAntonio, L. and Eisenberg D. (1989a). Hydrophobic Organisation Of Membrane Proteins. *Science.* **245,** 510-513.

Rees, D. C., Komiya, H., Yeates, T. O., Allen, J. P. and Feher, G. (1989b). The Bacterial Photosynthetic Reaction Center As A Model For Membrane Proteins. *Annu. Rev. Biochem.* **58,** 607-633.

Rees, D. C. (1990). Three-Dimensional Protein Structure Prediction Workshop: Overview and Summary. *Current Research in Protein Chemistry: Techniques, Structure, and Function.* Editor: Villafranca, J.J. Academic Press, London - Published under the Auspices of the Protein Society. Section VI - Structure Prediction Workshop. Page 551-556. ISBN 0-12-721956-0.

Regan, J. W., Kobilka, T. S., Yangfeng, T. L., Caron, M. G., Lefkowitz, R. J. and Kobilka, B. K. (1988). Cloning And Expression Of A Human-Kidney cDNA For An Alpha-2-Adrenergic Receptor Sub-type. *Proc. Natl. Acad. Sci. USA.* **85(17),** 6301-6305.

Reid, K. S. C., Lindley, P. F. and Thornton, J. M. (1985). Sulphur-aromatic interactions in proteins. *FEBS.* **190,** 2, 209-213.

Reid, L. S. and Thornton, J. M. (1989). Rebuilding Flavodoxin From C^{α} Co-ordinates: *A Test Study. Proteins: Structure, Function, and Genetics.* **5,** 170-182.

Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* **34,** 167-339.

Richardson, J. S. and Richardson, D. C. (1989). Principles and Patterns of Protein Conformation. In: *Prediction of Protein Structure and the Principles of Protein Conformation.* Editor: Fasman, G. D. Plenum Press, New York. ISBN: 0-306-43131-9. Chapter 1, 1-98.

Riegler, R. (1990). "Molecular recognition, allosteric receptors and drug design". Molecular structure, dynamics and interactions in bioactive membrane peptides. *Round Table Roussel-Uclaf.* **67,** 19-21.

Robson, B. and Pain, R. H. (1971). Analysis Of The Code Relating Sequence To Conformation In Proteins: Possible Implications For The Mechanisms Of Formation Of Helical Regions. *J. Mol. Biol.* **58,** 237-259.

Robson, B. and Suzuki, E. (1976). Conformational properties of amino acids residues in globular proteins. *J. Mol. Biol.* **107,** 327-356.

Rojewska, D. and Elber, R. (1990). Molecular Dynamics Study of Secondary Structure Motions in Protein: Application to Myohemerythrin. *Proteins Structure, Function, and Genetics.* **7,** 265-279.

Ross, E. M. (1989). Signal Sorting And Amplification Through G-Protein Coupled Receptors. *Neurone.* **3(2),** 141-152.

Rost, B. and Sander, C. (1993). Prediction of Protein Secondary Structure at Better than 70% Accuracy. *J. Mol. Biol.* **232,** 584-599.

Rost, B. and Sander, C. (1994). Combining Evolutionary Information And Neural Networks To Predict Protein Secondary Structure. *Proteins Structure Function And Genetics.* **19(1),** 55-72.

Rost, B., Sander, C. and Schneider, R. (1994). Redefining the Goals of Protein Secondary Structure Prediction. *J. Mol. Biol.* **235,** 13-26.

Ryba, N. J. P., Hall, M. D., and Findlay, J. B. C. (1992). Rhodopsin. In: *Molecular Biology of G-Protein-Coupled Receptors.* Edited by Brann, M. R. Birkhauser. Birkhauser, Switzerland. ISBN: 3764334657. Chapter 1, pages 1-30.

Šali, A., Shakhnovich, E. and Karplus, M. (1994). How does a protein fold? *Nature,* **369,** 248-251.

Sankararamakrishnan, R. and Vishveshwara, S. (1990). Conformational Studies on Peptides with Proline in the Right-Handed α-Helical Region. Biopolymers. **30,** 287-298.

Sankararamakrishnan, R., Sreerama, N. and Vishveshwara, S. (1991) Characterization of proline-containing right-handed α-helix by molecular dynamics studies. Biophysical Chemistry. **40,** 97-108.

Saunders, J. and Findlay, B. C. (1990). Monoamine Neurotransmitter Receptors: Ligand-Receptor Models. In: Protein structure, prediciton an *design.* Editors: Kay, J., Lunt, G. G. and Osguthorpe, D. J. *Biochem Soc Symp.* **57,** 81-90. ISBN: 1 85578 002 X.

Schiffer and Edmundson. (1967). Use Of Helical Wheels To Represent The Structures Of Proteins And To Identify Segments With Helical Potential. *Biophys. J.* **7,** 121.

Schofield, P. R., Rhee, L. M. and Peralta, E. G. (1987). Primary Structure Of The Human Beta-Adrenergic Receptor Gene. *Nucleic Acids.* **15(8),** 3636-3636.

Schulteis, G. and Koob, G. (1994). Dark Side Of Drug Dependence. *Nature.* **371,** 108-109.

Schiffer, M. and Edmundson, A. B. (1967). Use of helical wheels to represent the structures of proteins and to identify segments with helical potential. *Biophy J.* **7,** p121.

Schiffer, M., Chang, C.-H. and Stevens, F. J. (1992). The functions of tryptophan residues in membrane proteins. *Prot. Eng.* **5(3),** 213-214.

Seeman, P. (1980). Brain Dopamine Receptors. *Pharmacol. Rev.,* **32(3),** 229-313.

Seeman, P., Ulpian, C., Grigoriadis, D., Pri-Bar, I. and Buchman O. (1985). Conversion Of Dopamine D_1 Receptors From High To Low Affinity For Dopamine. *Biochem. Pharmacol.* **34,** 151-154.

Seeman, P., Watanabe, M., Grigoriadis, D., Tedesco, J. L., George, S. R., Svensson, U., Nilsson, J. L. G. and Neumayer, J. L. (1985). Dopamine D_2 Receptor Binding Sites For Agonists - A Tetrahedral Model. *Mol. Pharmacol.* **28(5),** 391-399.

Seeman, P. (1987). Dopamine receptors and the dopamine hypothesis of schizophrenia. *Synapse.* **1,** 133-152.

Seeman, P., Bzowej, N. H., Guan, H. C., Bergeron C, Reynolds, G. P., Bird, E. D., Reiderer, P., Jellinger, K. and Tourtellotte, W. W. (1987). Human brain D1 and D2 dopamine receptors in schizophrenia, Alzheimer's, Parkinson's and Huntington's diseases. *Neuropsychopharmacology.* **1,** 5-15.

Seeman, P. and Niznik H. B. (1990). Dopamine receptors and transporters in Parkinson's disease and schizophrenia. *FASEB J.* **4,** 2737-2744.

Selbie, L. A., Hayes, G. and Shine, J. (1989). The Major D2 Receptor - Molecular Analysis Of The Human D2A Subtype. *DNA.* **8(9),** 683-689.

Shaikh, S., Collier, D., Kerwin, R. W., Pilowsky, L. S., Gill, M., Xy, W-M. and Thornton, A. (1993). Dopamine D_4 receptor subtypes and response to clozapine. *Lancet.* **341(Jan 9),** 116.

Shakhnovich, E. I. (1994). Proteins With Selected Sequences Fold Into Unique Native Conformations. *Phys. Rev. Lett.* **72(24),** 3907-3910.

Shaw, W. V. (1987). Protein Engineering - The Design, Synthesis and Characterization of Factitious Proteins. *Biochem J.* **246(1),** 1-17.

Shiver, J. W., Peterson, A. A., Widger, W. R., Furbacher, P. N. and Cramer, W. A. (1989). Prediction of Bilayer Spanning Domains of Hydrophobic and Amphipathic Membrane Proteins - Application to the Cytochrome-B and Colicin Families. *Meth. Enzymol.* **172,** 439-461.

Sibley, D. R., De Lean, A. and Creese, I. (1982). Anterior Pituitary Dopamine Receptors. Demonstration Of Interconvertible High And Low Affinity States Of The D_2 Dopamine Receptor. *J. Biol. Chem.* **257,** 6351-6361.

Sibley, D. R. and Monsma, F. J. (1992). Molecular Biology Of The Dopamine Receptors. *TIPS.* **13(2),** 61-69.

Singer, M. (1994). Personal communication - email from Michael Singer (mike@miris.med.yale.edu) to: 7tms_r@net.bio.net on Tuesday, 1st November, 1994.

Singer, S. J. (1971). In: Structure and Function of Biological Membranes (L. I. Rothfield, ed.). Academic Press, New York, 146-223.

Singer, S. J., and Nicolson, G. L. 1972. The fluid mosaic model of the structure of cell membranes. *Science.* **175,** 720.

Singleton, P. and Sainsbury, D. (1978). Dictionary of Microbiology. John Wiley & Sons, Chichester. 1st Edition. ISBN: 0 471 99658 0. Page 326.

Smith, S. O., Palings, I., Copie, V., Raleigh, D. P., Courtin, J., Pardoen, J. A., Lugtenburg, J., Mathies, R. A. and Griffin, R. G. (1987). Low-temperature solide-state ^{13}C NMR studies of the retinal chromophore in rhodopsin. *Biochem.* **26,** 1606-1611.

Smith, W. (1990). Molecular Dynamics on Hypercube Parallel Computers. In Allan, R.J. and Guest, M.F. (eds). *Intel iPSC/2 Workshop Proceedings.* Advanced Research Computing Group, Daresbury Laboratory, Warrington, UK. pp 1-11.

Sokoloff, P., Giros, B., Martres, M. P., Bouthenet, M. L. and Schwartz, J. C. (1990). Molecular Cloning And Characterization Of A Novel Dopamine Receptor (D3) As A Target For Neuroleptics. *Nature,* **347(6289),** 146-151. SWISS-PROT Fetch Number: P19020.

Sternberg, M. J. E. and Thornton, J. M. (1978). Prediction of protein structure from amino acid sequence. *Nature.* **271,** 15-20.

Strader, C. D., Candelore, M. R., Hill, W. S., Dixon, R. A. F. and Sigal, I. S. (1989a). Identification of two serine residues involved in agonist activation of the β-adrenergic receptor function. *J. Biol. Chem.* **264,** 13572-13578.

Strader, C. D., Sigal, I. S. and Dixon, R. A. F. (1989b). Genetic Approaches To The Determination Of Structure Function Relationships Of G Protein Coupled Receptors. *Trends Pharmacol. Sci. (Suppl.).* **(Dec), 10,** 26-30.

Strader, C. D., Candelore, M. R., Hill, W. S., Dixon, R. A. F. and Sigal, I. S. (1989c). A single amino acid substitution in the β-adrenergic receptor promotes partial agonist activity from antagonists. *J. Biol. Chem.* **264,** 16470-16477.

Strader, C. D. and Dixon, A. F. (1992). Genetic Analysis of the β-Adrenergic Receptor. In: *Molecular Biology of Receptors That Couple To G-Proteins. Applications of Molecular Genetics to Pharmacology.* Editor: Brann, M. R. Birkhauser, Switzerland. ISBN: 3764334657. Chapter 3. Pages 62-75.

Summers, N. L. and Karplus, M. (1989). Construction of Side-chains in Homology Modelling - Application to the C-terminal Lobe of Rhizopuspepsin. *J. Mol. Biol.* **210,** 785-811.

Sunahara, R. K., Guan, H. C., Odowd, B. F., Seeman, P., Laurier, L. G., Ng, G., George, S. R., Torchia, J., Vantol, H. H. M. and Niznik, H. B. (1991). Cloning Of The Gene For A Human Dopamine D5-Receptor With Higher Affinity For Dopamine Than D1. *Nature.* **350(6319),** 614-619. SWISS-PROT Fetch Number: P21918.

Sussman, J. L., Harel, M., Frolow, F., Oefner, C., Goldman, A., Toker, L. and Silman, I. (1991). Atomic Structure of Acetylcholinesterase from *Torpedo californica:* A Prototypic Acetylcholine-Binding Protein. *Science.* **253,** 872-879.

Sutcliffe, M. J., Hayes, F. R. F. and Blundell, T. L. (1987). Knowledge Based Modelling Of Homologous Proteins, Part II: Rules For The Conformations Of Substituted Side-Chains. *Protein Engineering.* **1(5),** 385-392.

Suzuki, M. (1992). DNA-Bridging By A Palindromic α-Helix. *Proc. Natl. Acad. Sci. USA. 89,* 8726-8730.

Sylte, I., Edvardsen, Edvardsen, Ø. and Dahl, S. G. (1993). Molecular Dynamics Of The 5-HT_{1A} Receptor And Ligands. *Protein Eng.* **6(7),** 691-700.

Szent-Györgyi, A. G. and Cohen, C. (1957). Role Of Proline In Polypeptide Chain Configurations Of Proteins. *Science.* **126,** 697.

Tabak, D. (1990). *Multiprocessors.* Prentice-Hall International, London,

Tomita, M., and Marchesi, V. T. (1975). Amino-acid sequence and oligosaccharide attachment sites of human erythrocyte glycophorin. *Proc. Natl. Acad. Sci. USA.* **72,** 2964.

Tregidgo R.W.S. and Downton, A.C. (1990). Processor farm analysis and simulation for embedded parallel processing systems. *Proceedings of the 11th Occam User Group Technical Meeting.* IOS Press, Amsterdam.

Tripos Inc. Subsidiary of Evans & Sutherland, 6548 Clayton Road, St. Louis, MO 62117. Supplier of SYBYL molecular modelling software and Evans & Sutherland 10/33 workstation.

Trumpp-Kallmeyer, S., Hoflack, J., Bruinvels, A. and Hibert, M. (1992). Modelling of G-Protein-Coupled Receptors: Application to Dopamine, Adrenaline, Serotonin, Acetylcholine, and Mammalian Opsin Receptors. *J. Med. Chem.* **35,** 3448-3462.

Tüchsen, E. and Woodward, C. (1987). Assignment of Asparagine-44 Side-Chain Primary Amide ^{1}H NMR Resonances and the Peptide Amide N^{1}H Resonance of Glycine-37 in Basic Pancreatic Trypsin Inhibitor. *Biochemistry.* **26,** 1918-1925.

Urwyler, S. and Markstein, R. (1986). Identification Of Dopamine D_3 And D_4 Binding - Sites, Labelled With [H-3]2-Amino-6,7-Dihydroxy-1,2,3,4-Tetrahydronaphthalene, As High Agonist Affinity States Of The D_1 And D_2 Dopamine Receptors Respectively. *J. Neurochem.* **46(4),** 1508-1067.

Van Drie, J. H., Weininger, D. and Martin, Y. (1989). ALLADIN: An integrated tool for computer-assisted molecular design and pharmacophore recognition from geometric, steric and substructure searching of three-dimensional molecular structures. *J. Comp.Aided Mol. Des.* **3,** 225-251.

Van Günsteren, W. F., Berendsen, H. J. C., Hermans, J., Hol, W. G. J. and Postma, J. P. M. (1983). Computer simulation of the dynamics of hydrated protein crystals and its comparison with X-ray data. *Pro. Natl. Acad. Sci. USA.* **80(4),** 4315-4319.

Van Oene, J. C. and Horn, A. S. (1985). The pharmacological characterisation of 3,4-dihydroxyphenylimino-2-imidazolidine (DPI) as a potent mixed α_1/α_2-adrenoceptor agonist rather than as a dopamine receptor agonist. *J. Pharmacy and Pharmacology.* **37(11),** 844-847.

Van Tol, H. H. M., Bunzow, J. R., Guan, H. C., Sunahara, R. K. and Niznik, H. B. (1991). Cloning Of The Gene For A Human Dopamine D4-Receptor With High-Affinity For The Antipsychotic Clozapine. *Nature.* **350(6319),** 610-614. SWISS-PROT Fetch Number: P21917.

Van Tol, H. H. M., Wu, C. M., Guan, H., Ohara, K., Bunzow, J. R., Civelli, O., Kennedy, Seeman, P., Niznik, H. B. and Jovanovic, V. (1992). Multiple Dopamine D_4 Receptor Variants In The Human Population. *Nature.* **358,** 149-152.

Verdonk, M. L., Boks, G. J., Kooijman, H., Kanters, J. A. and Kroon, J. (1993). Sterochemistry of charged nitrogen-aromatic interactions and its involvement in ligand-receptor binding. *J Comp Aided Mol Design.* **7(2),** 173-182.

Verma, M. (1986). Molecular Cloning And Sequencing Of Lysozyme Gene Of Bacteriophage SF6 of *Bacillus subtilus. Curr. Microbiol.* **13,** 299-301.

Vlijmen, H. W. Th. van, IJzermann, A. P. (1989). Molecular Modeling Of A Putative Antagonist Binding Site On Hellix III Of The β-adrenoceptor. *J. Computer-Aided Mol. Design.* **3,** 165-174.

Vogt, G. and Argos, P. (1992). Searching for distantly related protein sequences in large databases by parallel processing on a transputer machine. *Comput. Applic. Bio.,* **8,** 49-55.

Von Heijne, G. (1991). Computer analysis of DNA and protein sequences. *Eur. J. Biochem.,* **199,** 253-256.

Von Heijne, G. (1991). Proline Kinks In Transmembrane α-Helices. *J. Mol. Biol.* **218,** 499-503.

Wallace, B. A., Cascico, M. and Mielke, D. L. (1986). Evaluation of methods for the prediction of membrane protein secondary structure. *Proc. Natl. Acad. Sci. U.S.A.* **83,** 9423-9427.

Waters, J. A., Spivak, C. E., Hermsmeir, M., Yadav, J. S., Liang, R. F. and Gund, T. M. (1988). Synthesis, Pharmacology, and Molecular Modelling Studies of Semi-rigid, Nicotinic Agonists. *J. Med. Chem.* **31(3),** 545-554.

Watson, S. and Arkinstall, S. (1994). THE G-PROTEIN LINKED RECEPTOR *Facts Book.* Academic Press Ltd., London. ISBN 0-12-738440-5

Wang, H. Y., Lipfert, L., Malbon, C. C. and Bahouth, S. (1989). Site-directed antibodies define the topography of the β-adrenergic receptor. *J. Biol. Chem.* **264(24),** 14424-144431.

Weiner, S. J. and Kollman, P. A. (1981). AMBER: Assisted Model Building with Energy Refinement. A General Program for Modeling Molecules and Their Interactions. *J. Comp. Chem.* **2(3),** 287-303.

Weiner, S. J., Kollman, P. A., Case, D. A., Singh U. C.. Ghio, C. Alagona, G., Profeta, S. and Weiner P. (1984). A New Force Field for Mechanical Mechanical Simulation of Nucleic Acids and Proteins. *J. Am. Chem. Soc.* **106(3),** 765-784.

Weiner, J. S., Kollman, P. A., Nguyen, D. T. and Case, D. A. (1986). An All Atom Force Field for Simulations of Proteins and Nucleic Acids. *J. Comp. Chem.* **7(2),** 230-252.

Weis, M. S. and Schulz, G. E. (1993). Porin Conformation In The Absence Of Calcium - Refined Structure At 2 Angstrom Resolution. *J. Mol. Biol.* **231(3),** 817-824.

Wess, J., Buhl, T., Lambrecht, G. and Mutschler, E. (1990). Cholinergic Receptors. In: *COMPREHENSIVE MEDICINAL CHEMISTRY - The Rational Design, Mechanistic Study & Therapeutic Application of Chemical Compounds.* Editors: Hansch, C., Sammes, P. G. and Taylor, J. B. Pergamon Press, Oxford. **3;** Chapter 12.6; page 448.

West, R. (1991). PARKINSON'S DISEASE. Pamphlet published by the Office of Health Economics. ISSN 0473 8837.

Wetlaufer, D. B. (1973). Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins. *Proc. Nat. Acad. Sci. U.S.A.* **70(3),** 697-701.

Williams, D. H., Searle, M. S., Mackay, J. P., Gerhard, U. and Maplestone, R. A. (1993). Toward An Estimation Of Binding Constants In Aqueous Solution Studies Of Associations Of Vanomycin Group Antiobiotics. *Proc. Natl. Acad. Sci. U.S.A.* **90(4),** 1172-1178.

Williams, J., Elleman, T. C., Kingston, I. B., Wilkins, A. G. and Kuhn, K. A. (1982). The Primary Structure Of Hen Ovotransferrin. *Eur. J. Biochem.* **122(2),** 297-303.

Williams, K. A. and Deber, C. M. (1991). Proline Residues in Transmembrane Helices: Structural or Dynamic Role. *Biochem.* **30,** 8919-8923.

Williams, K. A. and Deber, C. M. (1993). Mutagenesis Of Bacteriophage Ike Major Coat Protein Transmembrane Domain: Role Of An Interfacial Proline Residue. *Biochem. Biophysical Res. Comm.* **196(1),** 1-6.

Williamson, M. P. (1994). The Structure And Function Of Proline-Rich Regions In Proteins. *Biochem. J.* **297,** 249-260.

Wilmot, C. M. and Thornton, J. M. (1988). Analysis and Prediction of the Different Types of β-turn in Proteins. *J. Mol. Biol.* **203(1),** 221-232.

Wishart, D. S. and Muir, A. K. (1990). Protein Structure Prediction. *Current Research in Protein Chemistry: Techniques, Structure, and Function.* Editor: Villafranca, J.J. Academic Press, London - Published under the Auspices of the Protein Society. Section VI - Structure Prediction Workshop. Page 557-565. ISBN 0-12-721956-0.

Wong, L. S., Eshel, G., Dreher, J., Ong, J and Jackson, D. M. (1991). Role of dopamine and GABA in the control of motor activity elicited from the rat nucleus accumbens. *Pharmacol. Biochem. Behav.* **38(4),** 829-885.

Wong, A. O., Chang, J. P. and Peter, R. E. (1992). Dopamine Stimulates Growth Hormone Release From The Pituitary of Goldfish, *Carassius auratus,* Through The Dopamine D_1 Receptors. *Endocrinology,* **130(3),** 1201-1210.

Wong, A. O., Chang, J. P. and Peter, R. E. (1993a). *In vitro* and *in vivo* evidence that dopamine exerts growth hormone-releasing activity in goldfish. *Am. J. Physiol.* **264(6),** 925-932.

Wong, A. O., Chang, J. P. and Peter, R. E. (1993b). Characterization Of D_1 Receptors Mediating Dopamine Stimulated Growth Hormone Release From Pituitary Cells Of The Goldfish, *Carassius auratus. Endocrinology,* **133(2),** 577-584.

Wong, A. O., Chang, J. P. and Peter, R. E. (1993c). Dopamine Functions As A Growth Hormone Releasing Factor In The Goldfish, *Carassius auratus. Fish Physiology and Biochemistry.* **11(1-6),** 77-84.

Wreggett, K. A. and De Lean. (1984). The Ternary Complex Model. Its Properties And Applications To Ligand Interactions With The D_2 Dopamine Receptor Of The Anterior Pituitary Gland. *Mol. Pharmacol.* **25,** 10-17.

Yeates, T. O., Kimiya, H., Rees, D. C., Allen, J. P. and Feher, G. (1987). Structure Of The Reaction Center From *Rhodobacter sphaeroides* R-26 - Membrane-Protein Interactions. *Proc. Natl. Acad. Sci. USA.* **84(18),** 6438-6442.

Zhou, W., Flanagan, C., Ballesteros, J. A., Konvicka, K., Davidson, J. S., Weinstein, H., Millar, R. P. and Sealfon, S. C. (1993). A Reciprocal Mutation Supports Helix 2 and Helix 7 Proximity In The Gonadotropin-Releasing Hormone Receptor. *Molecular Pharmacology.* **45(2),** 165-170.

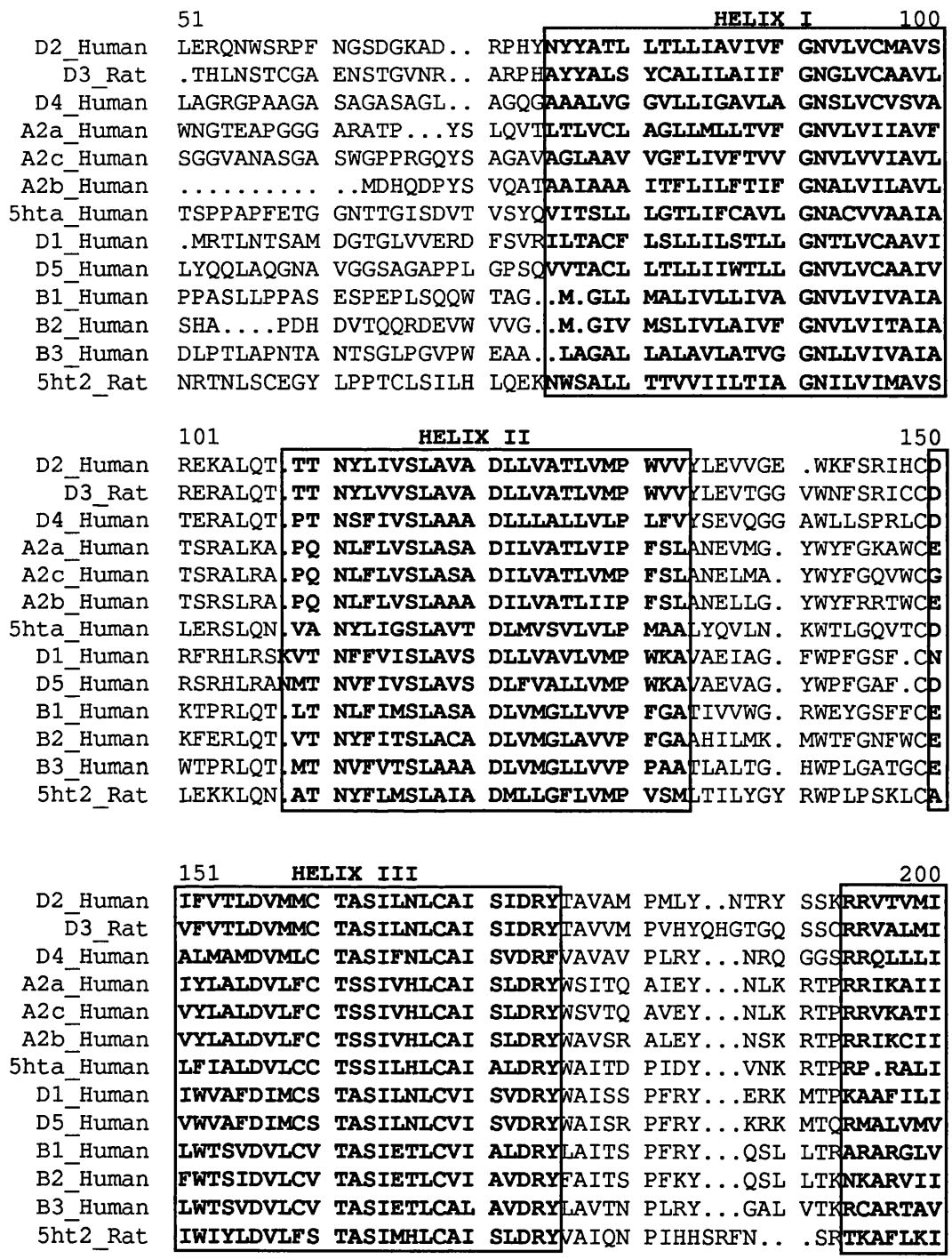
Zuckerkandl, E. and Pauling, L. (1965). In: *Evolving Genes and Proteins.* Editors: Bryson, V. and Hogel, H. J. Academic Press, New York. 97-166.

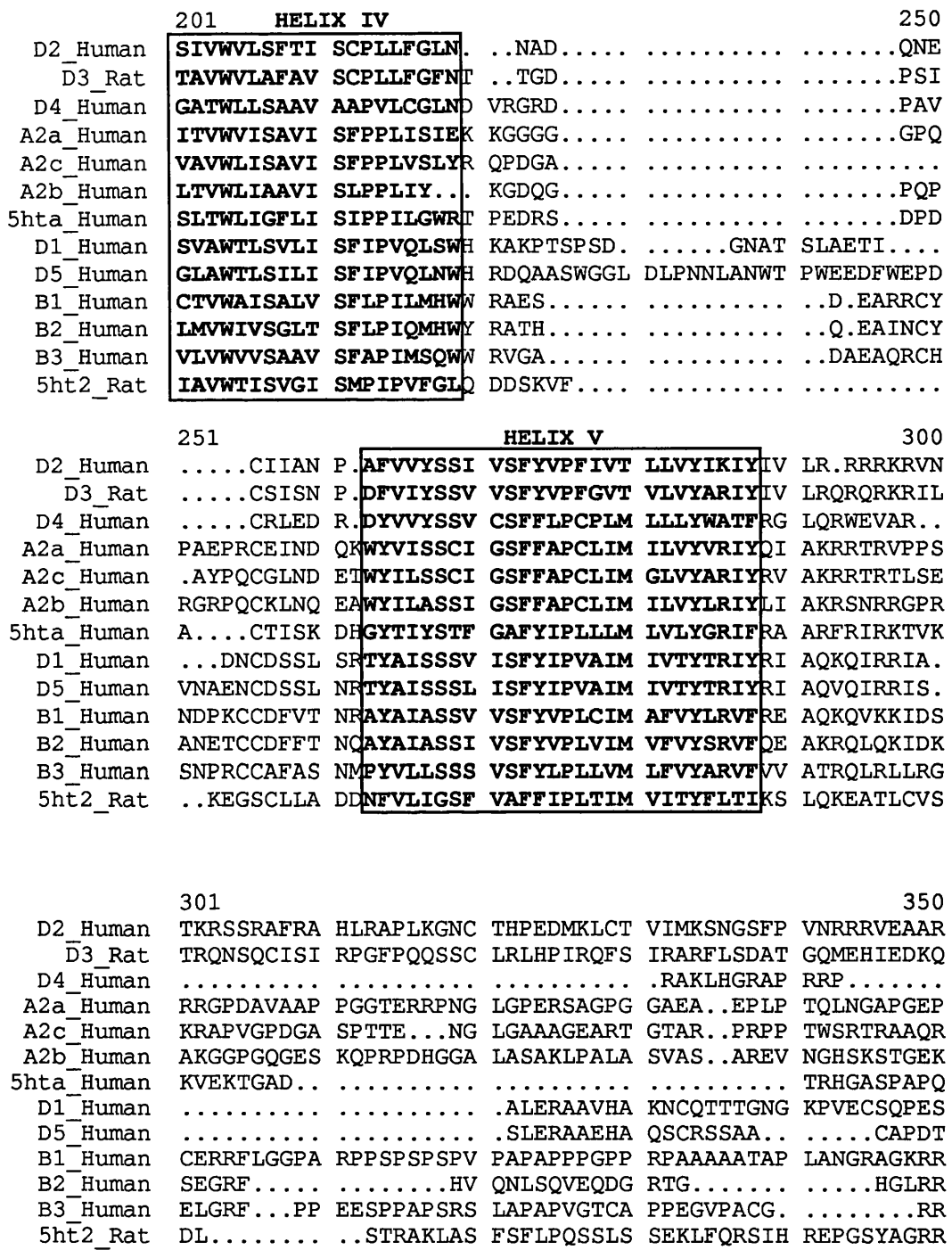
Zvelebil, M. M., Barton, G. J., Taylor, W. R. and Sternberg, M. J. E. (1987). Prediction of Protein Secondary Structure and Active Sites Using The Alignment of Homologous Sequences. *J. Mol. Biol.* **195(4),** 957-961.

APPENDICES

APPENDIX 1 Multiple Sequence File (MSF) created using PILEUP command (Devereux *et al.*, 1984). This MSF was used (as is) as input for Fourier Analysis of G protein coupled receptor (GPCR) Sequences using *Peppi!* coded by the author. The location of transmembrane helices are shown based on Baldwin's assignments for the dopamine sub-type GPCR D_1: transmembrane helix I (TM1): 75 to 100; TM2: 108 to 133; TM3: 150 to 175; TM4: 194 to 219; TM5: 263 to 288; TM6: 463 to 488; TM7: 508 to 533.

```
PileUp of: @paper3.fil

 Symbol comparison table: GenRunData:PileUpPep.Cmp  CompCheck:
1254

                  GapWeight: 3.0
             GapLengthWeight: 0.1

 Pileup.Msf  MSF: 654  Type: P  March 2, 1993  16:55  Check: 323
..

 Name: D2_Human          Len:   654  Check: 5030  Weight:  1.00
 Name: D3_Rat            Len:   654  Check: 4033  Weight:  1.00
 Name: D4_Human          Len:   654  Check: 9972  Weight:  1.00
 Name: A2a_Human         Len:   654  Check: 7559  Weight:  1.00
 Name: A2c_Human         Len:   654  Check: 5846  Weight:  1.00
 Name: A2b_Human         Len:   654  Check: 9510  Weight:  1.00
 Name: 5hta_Human        Len:   654  Check: 5723  Weight:  1.00
 Name: D1_Human          Len:   654  Check:  344  Weight:  1.00
 Name: D5_Human          Len:   654  Check: 9491  Weight:  1.00
 Name: B1_Human          Len:   654  Check: 9493  Weight:  1.00
 Name: B2_Human          Len:   654  Check: 8452  Weight:  1.00
 Name: B3_Human          Len:   654  Check: 5430  Weight:  1.00
 Name: 5ht2_Rat          Len:   654  Check: 9440  Weight:  1.00

//

            1                                                  50
  D2_Human  ..........  ..........  ..........  ........MD PLNLSWYDDD
    D3_Rat  ..........  ..........  ..........  ........MA PLSQIS....
  D4_Human  ..........  ..........  ..........  ........MG NRSTADADGL
 A2a_Human  ..........  ..........  ..........  ........MG SLQPDAGNAS
 A2c_Human  ..........  ..........  ...MASPALA AALAVAAAAG PNASGAGERG
 A2b_Human  ..........  ..........  ..........  .......... ..........
5hta_Human  ..........  ..........  ..........  ........MD VLSPGQGNNT
  D1_Human  ..........  ..........  ..........  .......... ..........
  D5_Human  ..........  ..........  ..........  ....MLPPGS NGTAYPGQFA
  B1_Human  ..........  ...MGAGVLV LGASEPGNLS SAAPLPDGAA TAARLLVPAS
  B2_Human  ..........  ..........  ..........  ....MGQPGN GSAFLLAPNR
  B3_Human  ..........  ..........  ..........  .MAPWPHENS S....LAPWP
  5ht2_Rat  MEILCEDNIS LSSIPNSLMQ LGDGPRLYHN DFNSRDANTS EASNWTIDAE
```


51
HELIX I
100
D2_Human LERQNWSRPF NGSDGKAD.. RPHYNYYATL LTLLIAVIVF GNVLVCMAVS
D3_Rat .THLNSTCGA ENSTGVNR.. ARPHAYYALS YCALILAIIF GNGLVCAAVL
D4_Human LAGRGPAAGA SAGASAGL.. AGQGAAALVG GVLLIGAVLA GNSLVCVSVA
A2a_Human WNGTEAPGGG ARATP...YS LQVTLTLVCL AGLLMLLTVF GNVLVIIAVF
A2c_Human SGGVANASGA SWGPPRGQYS AGAVAGLAAV VGFLIVFTVV GNVLVVIAVL
A2b_Human ....MDHQDPYS VQATAAIAAA ITFLILFTIF GNALVILAVL
5hta_Human TSPPAPFETG GNTTGISDVT VSYQVITSLL LGTLIFCAVL GNACVVAAIA
D1_Human .MRTLNTSAM DGTGLVVERD FSVRILTACF LSLLILSTLL GNTLVCAAVI
D5_Human LYQQLAQGNA VGGSAGAPPL GPSQVVTACL LTLLIIWTLL GNVLVCAAIV
B1_Human PPASLLPPAS ESPEPLSQQW TAG..M.GLL MALIVLLIVA GNVLVIVAIA
B2_Human SHA....PDH DVTQQRDEVW VVG..M.GIV MSLIVLAIVF GNVLVITAIA
B3_Human DLPTLAPNTA NTSGLPGVPW EAA..LAGAL LALAVLATVG GNLLVIVAIA
5ht2_Rat NRTNLSCEGY LPPTCLSILH LQEKNWSALL TTVVIILTIA GNILVIMAVS
101
HELIX II
150
D2_Human REKALQT.TT NYLIVSLAVA DLLVATLVMP WVVYLEVVGE .WKFSRIHCD
D3_Rat RERALQT.TT NYLVVSLAVA DLLVATLVMP WVVYLEVTGG VWNFSRICCD
D4_Human TERALQT.PT NSFIVSLAAA DLLLALLVLP LFVYSEVQGG AWLLSPRLCD
A2a_Human TSRALKA.PQ NLFLVSLASA DILVATLVIP FSLANEVMG. YWYFGKAWCE
A2c_Human TSRALRA.PQ NLFLVSLASA DILVATLVMP FSLANELMA. YWYFGQVWCG
A2b_Human TSRSLRA.PQ NLFLVSLAAA DILVATLIIP FSLANELLG. YWYFRRTWCE
5hta_Human LERSLQN.VA NYLIGSLAVT DLMVSVLVLP MAALYQVLN. KWTLGQVTCD
D1_Human RFRHLRSKVT NFFVISLAVS DLLVAVLVMP WKAVAEIAG. FWPFGSF.CN
D5_Human RSRHLRANMT NVFIVSLAVS DLFVALLVMP WKAVAEVAG. YWPFGAF.CD
B1_Human KTPRLQT.LT NLFIMSLASA DLVMGLLVVP FGATIVVWG. RWEYGSFFCE
B2_Human KFERLQT.VT NYFITSLACA DLVMGLAVVP FGAAHILMK. MWTFGNFWCE
B3_Human WTPRLQT.MT NVFVTSLAAA DLVMGLLVVP PAATLALTG. HWPLGATGCE
5ht2_Rat LEKKLQN.AT NYFLMSLAIA DMLLGFLVMP VSMLTILYGY RWPLPSKLCA
151
HELIX III
200
D2_Human IFVTLDVMMC TASILNLCAI SIDRYTAVAM PMLY..NTRY SSKRRVTVMI
D3_Rat VFVTLDVMMC TASILNLCAI SIDRYTAVVM PVHYQHGTGQ SSCRRVALMI
D4_Human ALMAMDVMLC TASIFNLCAI SVDRFVAVAV PLRY...NRQ GGSRRQLLLI
A2a_Human IYLALDVLFC TSSIVHLCAI SLDRYWSITQ AIEY...NLK RTPRRIKAII
A2c_Human VYLALDVLFC TSSIVHLCAI SLDRYWSVTQ AVEY...NLK RTPRRVKATI
A2b_Human VYLALDVLFC TSSIVHLCAI SLDRYWAVSR ALEY...NSK RTPRRIKCII
5hta_Human LFIALDVLCC TSSILHLCAI ALDRYWAITD PIDY...VNK RTPRP.RALI
D1_Human IWVAFDIMCS TASILNLCVI SVDRYWAISS PFRY...ERK MTPKAAFILI
D5_Human VWVAFDIMCS TASILNLCVI SVDRYWAISR PFRY...KRK MTQRMALVMV
B1_Human LWTSVDVLCV TASIETLCVI ALDRYLAITS PFRY...QSL LTRARARGLV
B2_Human FWTSIDVLCV TASIETLCVI AVDRYFAITS PFKY...QSL LTKNKARVII
B3_Human LWTSVDVLCV TASIETLCAL AVDRYLAVTN PLRY...GAL VTKRCARTAV
5ht2_Rat IWIYLDVLFS TASIMHLCAI SLDRYVAIQN PIHHSRFN.. .SRTKAFLKI


201 HELIX IV 250
D2_Human SIVWVLSFTI SCPLLFGLN. ..NAD..... ....QNE
D3_Rat TAVWVLAFAV SCPLLFGFNT ..TGD..... ....PSI
D4_Human GATWLLSAAV AAPVLCGLND VRGRD..... ....PAV
A2a_Human ITVWVISAVI SFPPLISIEK KGGGG..... ....GPQ
A2c_Human VAVWLISAVI SFPPLVSLYR QPDGA..... ....
A2b_Human LTVWLIAAVI SLPPLIY... KGDQG..... ....PQP
5hta_Human SLTWLIGFLI SIPPILGWRT PEDRS..... ....DPD
D1_Human SVAWTLSVLI SFIPVQLSWH KAKPTSPSD. ......GNAT SLAETI....
D5_Human GLAWTLSILI SFIPVQLNWH RDQAASWGGL DLPNNLANWT PWEEDFWEPD
B1_Human CTVWAISALV SFLPILMHWW RAES..... ....D.EARRCY
B2_Human LMVWIVSGLT SFLPIQMHWY RATH..... ....Q.EAINCY
B3_Human VLVWVVSAAV SFAPIMSQWW RVGA..... ....DAEAQRCH
5ht2_Rat IAVWTISVGI SMPIPVFGLQ DDSKVF.... ....
251 HELIX V 300
D2_Human .....CIIAN P.AFVVYSSI VSFYVPFIVT LLVYIKIYIV LR.RRRKRVN
D3_Rat .....CSISN P.DFVIYSSV VSFYVPFGVT VLVYARIYIV LRQRQRKRIL
D4_Human .....CRLED R.DYVVYSSV CSFFLPCPLM LLLYWATFRG LQRWEVAR..
A2a_Human PAEPRCEIND QKWYVISSCI GSFFAPCLIM ILVYVRIYQI AKRRTRVPPS
A2c_Human .AYPQCGLND ETWYILSSCI GSFFAPCLIM GLVYARIYRV AKRRTRTLSE
A2b_Human RGRPQCKLNQ EAWYILASSI GSFFAPCLIM ILVYLRIYLI AKRSNRRGPR
5hta_Human A....CTISK DHGYTIYSTF GAFYIPLLLM LVLYGRIFRA ARFRIRKTVK
D1_Human ...DNCDSSL SRTYAISSSV ISFYIPVAIM IVTYTRIYRI AQKQIRRIA.
D5_Human VNAENCDSSL NRTYAISSSL ISFYIPVAIM IVTYTRIYRI AQVQIRRIS.
B1_Human NDPKCCDFVT NRAYAIASSV VSFYVPLCIM AFVYLRVFRE AQKQVKKIDS
B2_Human ANETCCDFFT NQAYAIASSI VSFYVPLVIM VFVYSRVFQE AKRQLQKIDK
B3_Human SNPRCCAFAS NMPYVLLSSS VSFYLPLLVM LFVYARVFVV ATRQLRLLRG
5ht2_Rat ..KEGSCLLA DDNFVLIGSF VAFFIPLTIM VITYFLTIKS LQKEATLCVS
301 350
D2_Human TKRSSRAFRA HLRAPLKGNC THPEDMKLCT VIMKSNGSFP VNRRRVEAAR
D3_Rat TRQNSQCISI RPGFPQQSSC LRLHPIRQFS IRARFLSDAT GQMEHIEDKQ
D4_Human .... .RAKLHGRAP RRP.....
A2a_Human RRGPDAVAAP PGGTERRPNG LGPERSAGPG GAEA..EPLP TQLNGAPGEP
A2c_Human KRAPVGPDGA SPTTE...NG LGAAAGEART GTAR..PRPP TWSRTRAAQR
A2b_Human AKGGPGQGES KQPRPDHGGA LASAKLPALA SVAS..AREV NGHSKSTGEK
5hta_Human KVEKTGAD.. .... TRHGASPAPQ
D1_Human .... .ALERAAVHA KNCQTTTGNG KPVECSQPES
D5_Human .... .SLERAAEHA QSCRSSAA.. .....CAPDT
B1_Human CERRFLGGPA RPPSPSPSPV PAPAPPPGPP RPAAAAATAP LANGRAGKRR
B2_Human SEGRF..... .....HV QNLSQVEQDG RTG..... .....HGLRR
B3_Human ELGRF...PP EESPPAPSRS LAPAPVGTCA PPEGVPACG. .....RR
5ht2_Rat DL..... ..STRAKLAS FSFLPQSSLS SEKLFQRSIH REPGSYAGRR

```
                351                                                     400
  D2_Human  RAQELEMEML SSTSPPERT. ......RYSP IPPSHHQLTL PDPSHHGLHS
    D3_Rat  YPQKCQDPLL SHLQPPSPGQ THGGLKRYYS IC...QDTAL RHPSLEG..G
  D4_Human  .......... SGPGPPSPT. .......... ..PPAPRLPQ DPCGPDCAPP
 A2a_Human  APAGPRDTDA LDLEESS... ..SSDHAERP PGPRRPERGP RGKGKARASQ
 A2c_Human  PRGGAPGP.. ...LRRG... ..GRRRAGAE GGAGGADGQG AGPGAAQSGA
 A2b_Human  EEGETPEDTG TRALPPSWAA LPNSGQGQKE GVCGASPEDE AEEEEEEEEE
5hta_Human  PKKSVNGESG SR........ ..NWRLGVES KAGGALCANG AVRQGDDGAA
  D1_Human  SF........ .......... .......... .......... ..........
  D5_Human  SL........ .......... .......... .......... ..........
  B1_Human  PS........ .......... .......... .......... ..........
  B2_Human  SS........ .......... .......... .......... ..........
  B3_Human  PA........ .......... .......... .......... ..........
  5ht2_Rat  TMQ....... .......... .......... .......... ..........

                401                                                     450
  D2_Human  TPDSPAKPEK NGHAKD.HPK IAKIFEIQTM PN.......G KTRTSLKTMS
    D3_Rat  AGMSPVERTR NSLSPTMAPK LS..LEVRKL SN.......G RLSTSLRLGP
  D4_Human  APGLPPDPCG SNCAPPDAVR .AAALPPQTP PQ.......T RRRRRAKITG
 A2a_Human  VKPGDSLRGA GRGRRGSGRR LQ....GRG. ...RSASGLP RRRAGAGGQ.
 A2c_Human  LTASRSPGPG GRLSRASSRS VEFFLSRRR. ...RARSSVC RRK...VAQ.
 A2b_Human  EEECEPQAVP VSPASACSPP LQQPQGSRVL ATLRGQVLLG RGVGAIGGQW
5hta_Human  LEVIEVHRVG NS.......K EHLPLPSEAG PTPCAPASFE RKNERNAEA.
  D1_Human  .......... .......... .......... .......... ..........
  D5_Human  .......... .......... .......... .......... ..........
  B1_Human  .......... .......... .......... .......... ..........
  B2_Human  .......... .......... .......... .......... ..........
  B3_Human  .......... .......... .......... .......... ..........
  5ht2_Rat  .......... .......... .......... .......... ..........

                451                HELIX VI                             500
  D2_Human  .RRKLSQQKE KKATQMLAIV LGVFIICWLP FFITHILNIH CD........
    D3_Rat  LQPRGVPLRE KKATQMVVIV LGAFIVCWLP FFLTHVLNTH CQ.A......
  D4_Human  .........RE RKAMRVLPVV VGAFLLCWTP FFVVHITQAL CP.A......
 A2a_Human  .......NRE KRFTFVLAVV IGVFVVCWFP FFFTYTLTAV ...G......
 A2c_Human  .......ARE KRFTFVLAVV MGVFVLCWFP FFFIYSLYGI CREA......
 A2b_Human  WRRRAHVTRE KRFTFVLAVV IGVFVLCWFP FFFSYSLGAI CPKH......
5hta_Human  .KRKMALARE RKTVKTLGII MGTFILCWLP FFIVALVLPF CESS......
  D1_Human  ...KMSFKRE TKVLKTLSVI MGVFVCCWLP FFILNCILPF CGSGETQP..
  D5_Human  ...RASIKKE TKVLKTLSVI MGVFVCCWLP FFILNCMVPF CSGHPEGPPA
  B1_Human  ...RLVALRE QKALKTLGII MGVFTLCWLP FFLANVVKAF .HREL.....
  B2_Human  ...KF.CLKE HKALKTLGII MGTFTLCWLP FFIVNIVHVI .QDNL.....
  B3_Human  ...RLLPLRE HRALCTLGLI MGTFTLCWLP FFLANVLRAL GGPSL.....
  5ht2_Rat  .....SISNE QKACKVLGIV FFLFVVMWCP FFITNIMAVI CKESCNE...

                501                HELIX VII                            550
  D2_Human  ..CNIPEVLY SAFTWLGYVN SAVNPIIYTT FNIEFRKAFL KILHC.....
    D3_Rat  ..CHVSPELY RATTWLGYVN SALNPVIYTT FNVEFRKAFL KILSC.....
  D4_Human  ..CSVPPRLV SAVTWLGYVN SALNPVIYTV FNAEFRNVFR KALRACC...
 A2a_Human  ..CSVPPTLF KFFFWFGYCN SSLNPVIYTI FNHDFRRAFK KILCRGDRKR
 A2c_Human  ..CQVPGPLF KFFFWIGYCN SSLNPVIYTV FNQDFRPSFK HILFRRRRRG
 A2b_Human  ..CKVPPGLF QFFFWIGYCN SSLNPVIYTI FNQDFRRAFR RILCRPWTQT
5hta_Human  ..CHMPTLLG AIINWLGYSN SLLNPVIYAY FNKDFQNAFK KIIKCNFCRQ
  D1_Human  ..FCIDSNTF DVFVWFGWAN SSLNPIIYA. FNADFRKAFS TLLGCYRLCP
  D5_Human  GFPCVSETTF DVFVWFGWAN SSLNPVIYA. FNADFQKVFA QLLGCSHFCS
  B1_Human  ....VPDRLF VFFNWLGYAN SAFNPIIYC. RSPDFRKAFQ GLLC...CAR
  B2_Human  ....IRKEVY ILLNWIGYVN SGFNPLIYC. RSPDFRIAFQ ELLCLRRSSL
  B3_Human  ....VPGPAF LALNWLGYAN SAFNPLIYC. RSPDFRSAFR RLLC..RCGR
  5ht2_Rat  ...NVIGALL NVFVWIGYLS SAVNPLVYTL FNKTYRSAFS RYIQCQYKEN
```

```
            551                                                       600
  D2_Human  .......... .......... .......... .......... ..........
    D3_Rat  .......... .......... .......... .......... ..........
  D4_Human  .......... .......... .......... .......... ..........
 A2a_Human  IV........ .......... .......... .......... ..........
 A2c_Human  FRQ....... .......... .......... .......... ..........
 A2b_Human  AW........ .......... .......... .......... ..........
5hta_Human  .......... .......... .......... .......... ..........
  D1_Human  ATNNAIETVS INNNGAAMFS SHHEPRGSIS KECNLVYLIP HAV..GSSED
  D5_Human  RT..PVETVN ISNE...LIS YNQDIVFHKE IAAAYIHMMP NAVTPGNREV
  B1_Human  RAARRRHATH GDRPRASGCL ARPGP..... .........P PSPGAASDDD
  B2_Human  KAYGNGYSSN GNTGEQSGY. .......... .......... ....HVEQEK
  B3_Human  RLPPEPCAAA RPALFPSGVP AARSS..... .........P AQPRLCQRLD
  5ht2_Rat  RKPLQLILVN TIPALAYKSS QLQVGQKKNS QEDAEQTVDD CSMVTLGKQQ

            601                                                       650
  D2_Human  .......... .......... .......... .......... ..........
    D3_Rat  .......... .......... .......... .......... ..........
  D4_Human  .......... .......... .......... .......... ..........
 A2a_Human  .......... .......... .......... .......... ..........
 A2c_Human  .......... .......... .......... .......... ..........
 A2b_Human  .......... .......... .......... .......... ..........
5hta_Human  .......... .......... .......... .......... ..........
  D1_Human  LKKEEAAGIA RPLEKL.... ...SPALSV. ILDYDTDVSL EKIQPITQNG
  D5_Human  DNDEEEGPFD RMFQIYQTSP DGDPVAESVW ELDCEGEISL DKITPFTPNG
  B1_Human  DDDVVGATPP ARLLEPWAGC NGGAAADSDS SLDEPCRPGF ASESKV....
  B2_Human  ENKLLCEDLP G..TEDFVGH QGTVPSD... NIDSQGRNCS TNDSLL....
  B3_Human  G......... .......... .......... .......... ..........
  5ht2_Rat  SEENCTDNIE TVNEKVSCV. .......... .......... ..........

            651
  D2_Human  ....
    D3_Rat  ....
  D4_Human  ....
 A2a_Human  ....
 A2c_Human  ....
 A2b_Human  ....
5hta_Human  ....
  D1_Human  QHPT
  D5_Human  FH..
  B1_Human  ....
  B2_Human  ....
  B3_Human  ....
  5ht2_Rat  ....
```

APPENDIX 2 Richard Henderson's replies with regard to the appropriateness of using bacteriorhodopsin as a structural template for modeling GPCRs.

>From rh15@uk.ac.cambridge.mrc-molecular-biology Thu Jan 14 12:48:41 1993
Received: from uk.ac.nsf by cardiff.ac.uk; Thu, 14 Jan 93 12:46:24 GMT
Received: from gray.computing-service-internal.cambridge.ac.uk
by sun3.nsfnet-relay.ac.uk with Internet SMTP
id <sg.17822-0@sun3.nsfnet-relay.ac.uk>;
Thu, 14 Jan 1993 12:46:42 +0000
Received: from uk.ac.cam.mrc-lmb.al by ppsw1.cam.ac.uk
with SMTP (PP-6.0) Cambridge as ppsw.cam.ac.uk
id <20199-0@ppsw1.cam.ac.uk>; Thu, 14 Jan 1993 12:46:30 +0000
Received: by al.mrc-lmb.cam.ac.uk (Concentrix-2800 3.0/Alliant-5.0) id AA22741;
Thu, 14 Jan 93 12:47:49 EST
Date: Thu, 14 Jan 93 12:47:49 EST
From: Henderson "R." <rh15@uk.ac.cambridge.mrc-molecular-biology>
Message-Id: <9301141247.AA22741@al.mrc-lmb.cam.ac.uk>
To: spxcw@uk.ac.cardiff.thor
Subject: dopamine
Status: RO

TO: Chris Wood
FROM: Richard Henderson
14-Jan-1993
Dear Chris,

Thanks for your message. We do have more accurate coordinates for bacteriorhodopsin, but I have not sent them out to anyone yet. I am still unhappy about several parts that need further work. Rather than pollute the world with multiple, undocumented and contradictory coordinates, we will have just two sets - preliminary coordinates (1BRD) and final ones after refinement.

However, I can say that the differences in the positions of the helix axes are imperceptible in the two sets. It will make no difference at all to modelling of dopamine receptor. We are also working on the G-protein coupled receptor family, and we already know that the positions and the angles of the helices are substantially different. The angles may differ by as much as 10-15 degrees, whereas any changes in our published bR coordinates will be less than 1 degree. So what you really need is a better starting model for the G-protein coupled receptor family, rather than anything better on bR.

Joyce Baldwin and Gebhard Schertler, working with me here, have each got a paper in press on aspects of the G-protein coupled receptor family. It might be useful if you had copies of their papers to read, since they will not be published for a month or two. I will mention your interest to them, and see whether they would be prepared to send you a preprint.

Yours sincerely, Richard Henderson.
From rh15@mrc-lmb.cam.ac.uk Thu Mar 24 19:22:45 1994
Received: from ppsw2.cam.ac.uk (pp@snow.csi.cam.ac.uk [131.111.12.55] sender rh15@mrc-lmb.cam.ac.uk) by rockall.cent.gla.ac.uk (8.6.7/UK-2.2/rockall) with SMTP id TAA07690 for <gacu57@udcf.gla.ac.uk>; Thu, 24 Mar 1994 19:22:42 GMT
Received: from uk.ac.cam.mrc-lmb.al by ppsw2.cam.ac.uk
with SMTP-CAM (PP-6.0) as ppsw.cam.ac.uk
id <21649-0@ppsw2.cam.ac.uk>; Thu, 24 Mar 1994 19:22:35 +0000

Received: by al.mrc-lmb.cam.ac.uk (Concentrix-2800 3.0/MDTG-Vevision: 1.3 •al.mrc-lmb.cam.ac.uk)
id AA15561; Thu, 24 Mar 94 19:24:30 EST
From: Henderson "R." <rh15@mrc-lmb.cam.ac.uk>
Date: Thu, 24 Mar 94 19:24:30 EST
Message-Id: <9403241924.AA15561@al.mrc-lmb.cam.ac.uk>
To: gacu57
Subject: receptors
Status: R

TO: Chris Wood
FROM: Richard Henderson
24-Mar-1994

Dear Chris,

I'm pleased you have made progress. Joyce Baldwin has just returned from a meeting in New Orleans and she found there are literally hundreds of people making models of G-protein coupled receptors. We cannot possibly read all the papers or keep up with what people are doing. Also, since our goal is to try to work out experimentally the actual structure, we don't want to spend much time readinng about work which is bound to be approximate, and may be completely in error. Please feel free to visit, but we would not wish to schedule a talk.

Concerning Hibert, he is trying to defend his use of the 1BRD coordinates in his modelling. As I explained, we think this is too far from reality to be useful. There may well be an overall net rotation of the entire molecule if you align the bacteriorhodopsin coordinates with the as-yet-unknown rhodopsin coordinates, but we think this will only be part of the story. Joyce has rotated bacteriorhodopsin map so that it is viewed from the best angle to simulate the rhodopsin projection, and there are still substantial differences in the helix positions between the two molecules of up to 10 Angstroms. Also which helices look untilted or parallel are different. We think there is no point in filling up the literature further with a discussion that will be fully resolved once there is a 3D model. Gebhard Schertler and Vinzenz Unger are working now to colect the data from tilted crystals needed to determine an approximate low resolution structure for rhodopsin. There will be a paper once that work is complete. It is unlikely they will pay any attention to the Hibert paper.

Yours sincerely, Richard Henderson.

Appendix 3

Source code listings and support files. These are located on the enclosed disc (inner sleeve on back cover of the thesis). For full description and explanation, please use textline editor to view README file.

APPENDIX 4 List of force field parameters (in fact copy of SYBYL log file) for energy minimizing of transmembrane region. Individual helices having been previously mininimized to 0.1 kcals/mol using Kollman all-atom forcefield.

```
###

### User chrisw does not have permission to use the "at" command on this
### machine. The batch job will be submitted to run as a background process.
### Contact your system manager to enable your account to use "at".
###
###
### Submitting job d3_jobx for background processing on machine type
### esv.
###

Batch application is starting - Thu Feb 17 03:36:21 GB  1994

Sybyl 6.03, Created Sept 16, 1993
Copyright (c) TRIPOS ASSOCIATES, INC. All rights reserved, 1983, 1993.
  This material contains confidential and proprietary information of TRIPOS
  Associates, Inc. Use of a copyright notice is precautionary only and does
  not imply publication or disclosure.
Molecular Silverware copyright (c) 1990 Rohm and Haas Company, All Rights
  Reserved.
Sybyl> setvar TAILOR!FORCE_FIELD!DERIVATIVE DER_52
Sybyl> setvar TAILOR!FORCE_FIELD!DIELECTRIC_CONSTANT 5.000000
Sybyl> setvar TAILOR!FORCE_FIELD!DIELECTRIC_FUNCTION constant
Sybyl> setvar TAILOR!FORCE_FIELD!HBOND_RAD_SCALING 1.000000
Sybyl> setvar TAILOR!FORCE_FIELD!NON_BONDED_CUTOFF 10.000000
Sybyl> setvar TAILOR!FORCE_FIELD!ONE_FOUR_SCALING 0.500000
Sybyl> setvar TAILOR!FORCE_FIELD!PARAMETER_SET Koll_united
Sybyl> setvar TAILOR!FORCE_FIELD!REVIEW_HS_AND_LPS REVIEW
Sybyl> setvar TAILOR!FORCE_FIELD!SCALE!ANGLE_BEND 1.0
Sybyl> setvar TAILOR!FORCE_FIELD!SCALE!BOND_STRETCH 1.0
Sybyl> setvar TAILOR!FORCE_FIELD!SCALE!ELECTROSTATICS 1.0
Sybyl> setvar TAILOR!FORCE_FIELD!SCALE!FIXED_ANGLE 1.0
Sybyl> setvar TAILOR!FORCE_FIELD!SCALE!FIXED_DISTANCE 1.0
Sybyl> setvar TAILOR!FORCE_FIELD!SCALE!FIXED_RANGE 1.0
Sybyl> setvar TAILOR!FORCE_FIELD!SCALE!FIXED_TORSION 1.0
Sybyl> setvar TAILOR!FORCE_FIELD!SCALE!HBOND_C 1.0
Sybyl> setvar TAILOR!FORCE_FIELD!SCALE!HBOND_D 1.0
Sybyl> setvar TAILOR!FORCE_FIELD!SCALE!IMPROPER_TORSION 1.0
Sybyl> setvar TAILOR!FORCE_FIELD!SCALE!OOP 1.0
Sybyl> setvar TAILOR!FORCE_FIELD!SCALE!TORSION 1.0
Sybyl> setvar TAILOR!FORCE_FIELD!SCALE!VDW_EPSILON 1.0
Sybyl> setvar TAILOR!FORCE_FIELD!SCALE!VDW_RADIUS 1.0
Sybyl> setvar TAILOR!MAXIMIN2!BATCH_CHECKPOINT_INTERVAL 0
Sybyl> setvar TAILOR!MAXIMIN2!COLOR_OPTION Force
Sybyl> setvar TAILOR!MAXIMIN2!GRAPHICS_UPDATE 1
Sybyl> setvar TAILOR!MAXIMIN2!LIST_TERMS NO
Sybyl> setvar TAILOR!MAXIMIN2!LS_ACCURACY 0.001000
Sybyl> setvar TAILOR!MAXIMIN2!LS_STEP_SIZE 0.001000
Sybyl> setvar TAILOR!MAXIMIN2!MAXIMUM_ITERATIONS 50
Sybyl> setvar TAILOR!MAXIMIN2!MAX_DISPLACEMENT 0.010000
Sybyl> setvar TAILOR!MAXIMIN2!MINIMIZATION_METHOD CONJUGATE_GRADIENT
Sybyl> setvar TAILOR!MAXIMIN2!MIN_ENERGY_CHANGE 0.100
Sybyl> setvar TAILOR!MAXIMIN2!NON_BONDED_RESET 10
```

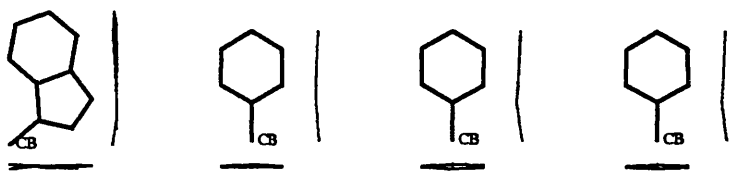
```
Sybyl> setvar TAILOR!MAXIMIN2!RESET_COUNT 100
Sybyl> setvar TAILOR!MAXIMIN2!RMS_DISPLACEMENT 0.001
Sybyl> setvar TAILOR!MAXIMIN2!RMS_GRADIENT 0.050
Sybyl> setvar TAILOR!MAXIMIN2!SIMPLEX_ITERATIONS 0
Sybyl> setvar TAILOR!MAXIMIN2!SIMPLEX_THRESHOLD 1000.000000
Sybyl> setvar TAILOR!MAXIMIN2!STATUS_UPDATE 1
Sybyl> setvar TAILOR!MAXIMIN2!TERMINATION_OPTION ENERGY_CHANGE
Sybyl> setvar TAILOR!MAXIMIN2!THRESHOLD -10.000000
Sybyl>
Sybyl>
Sybyl>
First pass . . .
NO_FILES_READ - No ASCII parameter files found in given directory

Ambiguous or unrecognized command "ATOM_DEF"
Sybyl>
First pass . . .
NO_FILES_READ - No ASCII parameter files found in given directory

Ambiguous or unrecognized command "BOND_DEF"
Sybyl>
Sybyl>   setvar MM_BATCH_INPUT_PAIRS  "M10 /usr2/people/chrisw/d3_jobx.mol2"
Sybyl>
Sybyl>   set autosave off
Sybyl>
Sybyl>   for m in  "M10 /usr2/people/chrisw/d3_jobx.min"
1>       mol in $m
1>   endfor
Reading molecule abcdefg_ok re: D3
Opening dictionary protein . . .
Sybyl>
Sybyl>   uims activate batch_mm_commands
Sybyl>
Sybyl>   BATCH_MIN /usr2/people/chrisw/d3_jobx.ff  NO
Field fit terms for abcdefg_ok re: D3 in M10
Sybyl>
Sybyl>   for m in  "M10 /usr2/people/chrisw/d3_jobx.mol2"
1>       mol out $m
1>   endfor
Sybyl>
Sybyl>   quit yes

Batch application is ending - Thu Feb 17 04:40:57 GB  1994

script.sh: Application completed.  Restarting NetBatch.
Connected to remote machine rhum ...
Connected to host machine rhum ...
Local TA_ROOT is "/usr/people/xray/sybyl60"
Retrieving files from remote machine rhum...
Batch job completed.  Batch directories will be removed.
```

Appendix 5

Distorted geometry output generated by Procheck (Laskowski *et al.*, 1993) with regard to dopamine receptor models (D_1 through to D_5) and muscarnic M_1 receptor model. For comparison, Procheck was applied to the β_2-AR model of Donnelly *et al.* (1994) and is included here.

Distorted geometry
D1-DAR

Planar groups

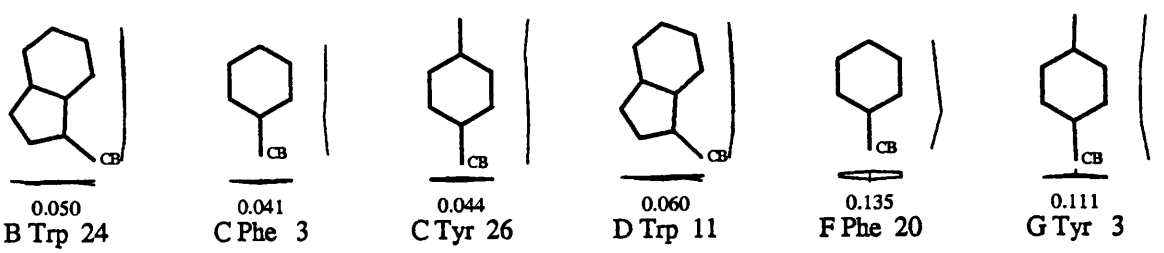


0.061 B Trp 24 | 0.046 F Phe 19 | 0.075 F Phe 20 | 0.065 G Phe 3

Sidechains with RMS dist. from planarity > 0.04A for rings, or > 0.03A otherwise. Value shown is RMS dist.

Distorted geometry
D2-DAR

Planar groups



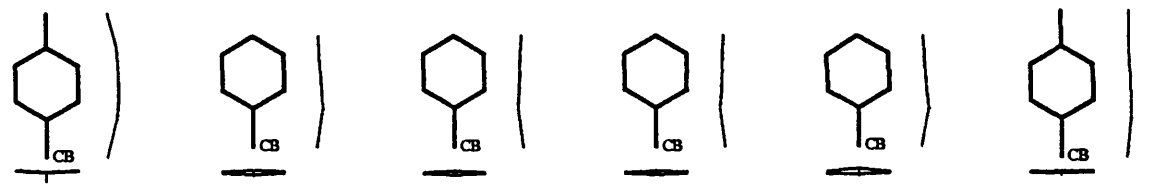
0.050 B Trp 24 | 0.041 C Phe 3 | 0.044 C Tyr 26 | 0.060 D Trp 11 | 0.135 F Phe 20 | 0.111 G Tyr 3

0.045 G Phe 6 | 0.139 G Tyr 11

Sidechains with RMS dist. from planarity > 0.04A for rings, or > 0.03A otherwise. Value shown is RMS dist.

Distorted geometry
D3-DAR

Planar groups

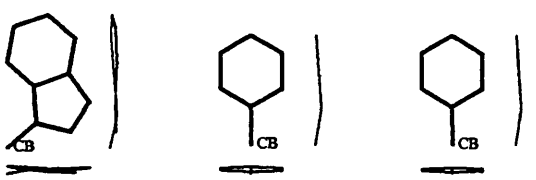


0.169	0.072	0.062	0.060	0.088	0.052
A Tyr 7	D Phe 25	E Phe 11	F Phe 12	F Phe 20	G Tyr 3

Sidechains with RMS dist. from planarity > 0.04A for rings, or > 0.03A otherwise. Value shown is RMS dist.

Distorted geometry
D4-DAR

Planar groups

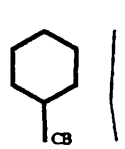


0.093	0.085	0.075
D Trp 11	F Phe 19	F Phe 20

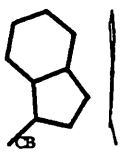
Sidechains with RMS dist. from planarity > 0.04A for rings, or > 0.03A otherwise. Value shown is RMS dist.

Distorted geometry
D5-DAR

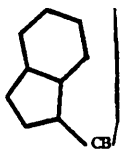
Planar groups



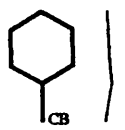
0.069
B Phe 6



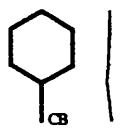
0.074
C Trp 3



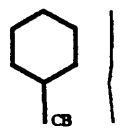
0.052
D Trp 11



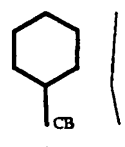
0.073
F Phe 19



0.057
F Phe 20



0.046
G Phe 3

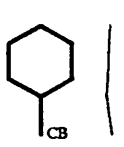


0.091
G Phe 6

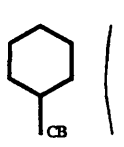
Sidechains with RMS dist. from planarity > 0.04A for rings, or > 0.03A otherwise. Value shown is RMS dist.

Distorted geometry
M1-GPCR

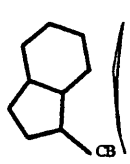
Planar groups



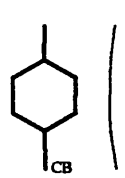
0.060
A Phe 25



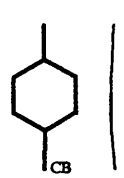
0.095
B Phe 20



0.139
C Trp 3



0.095
C Tyr 8

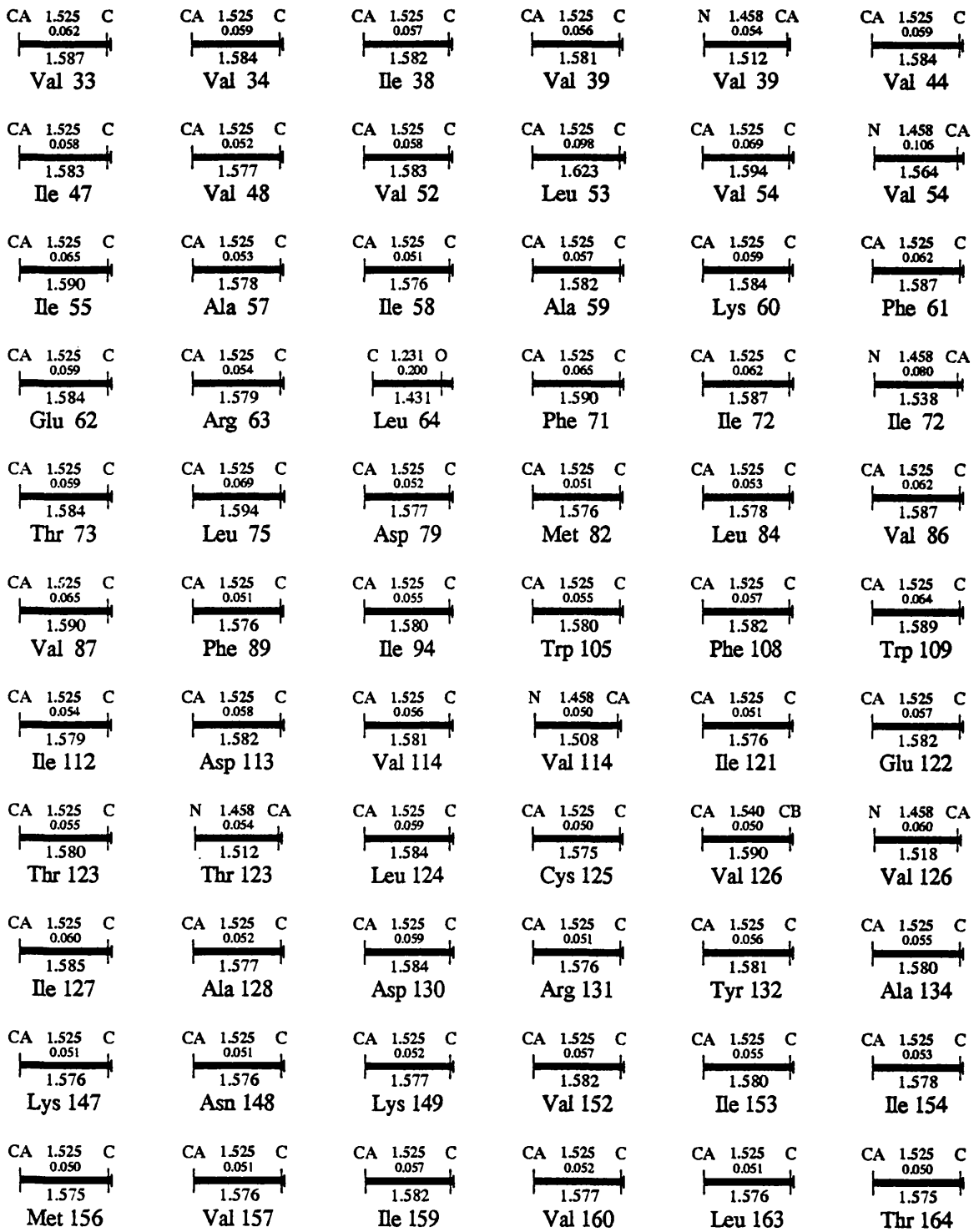


0.054
F Tyr 19

Sidechains with RMS dist. from planarity > 0.04A for rings, or > 0.03A otherwise. Value shown is RMS dist.

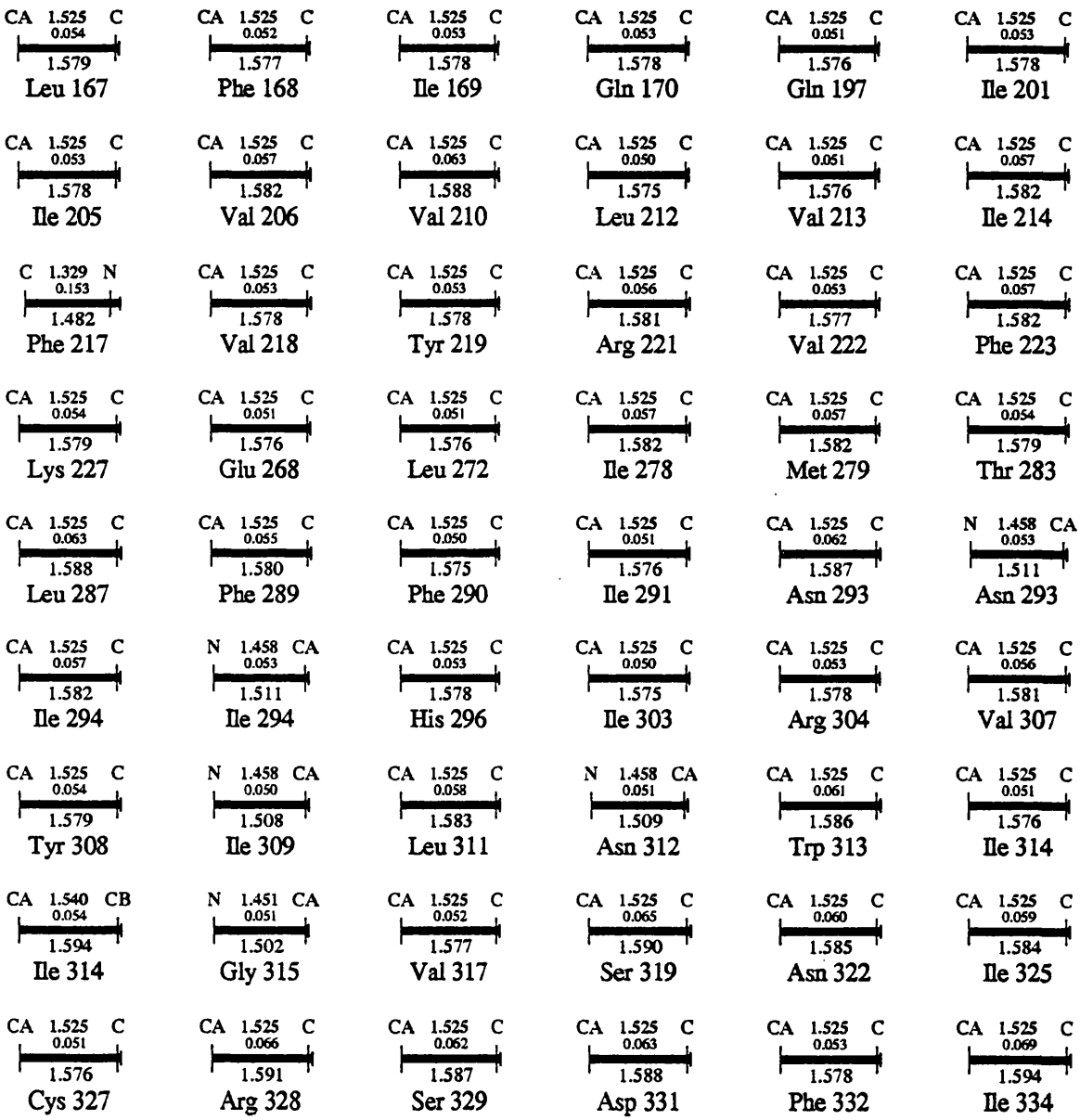
Distorted geometry
b2-AR

Main-chain bond lengths



Distorted geometry
b2-AR

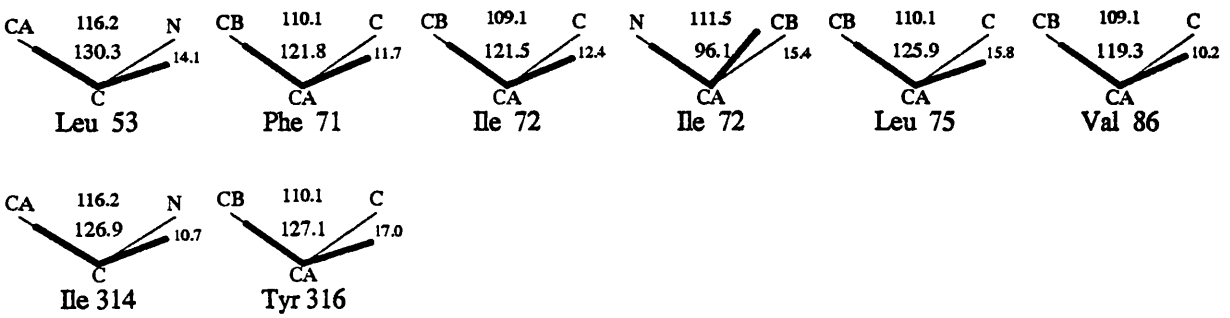
Main-chain bond lengths (contd)



Bonds differing by > 0.05A from small-molecule values. Values shown: "ideal", difference, actual

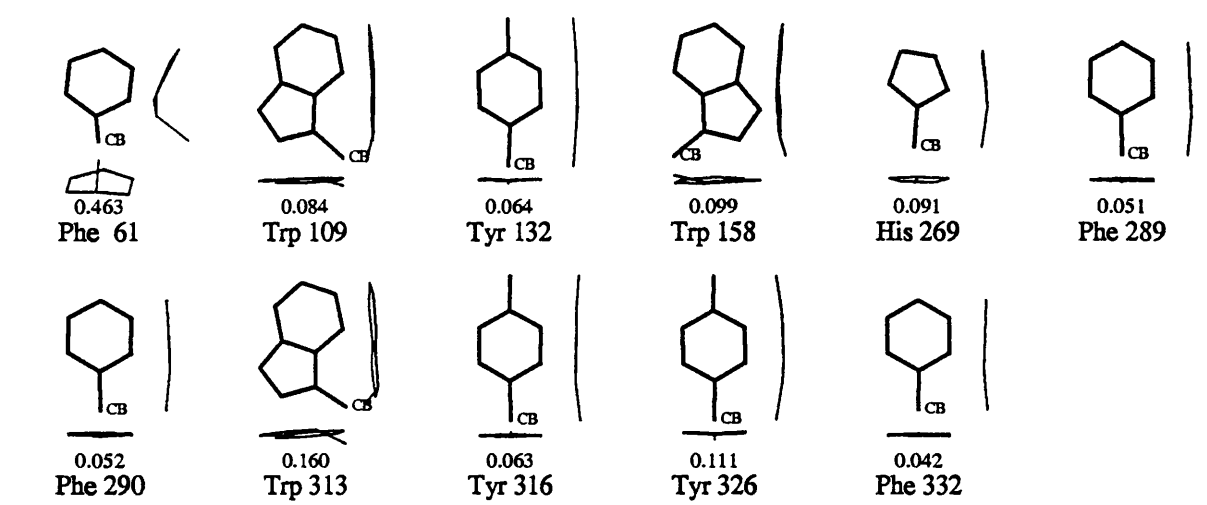
Distorted geometry
b2-AR

Main-chain bond angles



Bond angles differing by > 10.0 degrees from small-molec values. Values shown: "ideal", actual, diff.

Planar groups



Sidechains with RMS dist. from planarity > 0.04A for rings, or > 0.03A otherwise. Value shown is RMS dist.