STRUCTURE PREDICTION OF *A. thaliana* G PROTEIN ALPHA SUBUNIT USING BIOINFORMATICS TOOLS BASED ON SEQUENCE ALIGNMENTS

by
MERT ŞAHİN

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfillment of
the requirements for the degree of
Master of Science

Sabancı University
September 2002

STRUCTURE PREDICTION OF *A. thaliana* G PROTEIN ALPHA SUBUNIT USING BIOINFORMATICS TOOLS BASED ON SEQUENCE ALIGNMENTS

APPROVED BY:

Assoc. Prof. Zehra Sayers …….
(Dissertation Supervisor)

Assist. Prof. Uğur Sezerman …….
(Co-Supervisor)

Prof. Hüveyda Başağa …….

Assoc. Prof. Canan Baysal …….

Assist. Prof. Damla Bilgin …….

DATE OF APPROVAL: …….

ABSTRACT

Identification and characterization of plant G-proteins are mainly based on sequence homology with their mammalian counterparts. Studies with mutant plants have confirmed the assigned functional roles of these proteins and possible signaling pathways involving G-proteins (Fujisawa Y *et al.*, 2001). The work presented in this thesis, reports detailed sequence alignment analyses of *A. thaliana* G-protein α subunit (GPA1) with its mammalian counterpart rat transducin α subunit (PDB entry 1gg2). Alignment of GPA1 with the other known Gα sequences, from plants, reveal a high degree of homology. Same analyses are extended to *A. thaliana* β and γ subunits.

Results of sequence analyses were used as a basis for secondary structure prediction and for modeling 3D structure of GPA1. Secondary structure prediction was carried out for β and γ subunits as well.

Modeling calculations were based on the sequence alignment, secondary structure prediction, known structure of rat transducin α subunit, and carried out using the MODELLER module within insightII program. The best model predicted 76% correct folding within confidence limits determined using ERRAT. The accuracy of this model can be improved with further optimization of the loop regions. Further work is in progress for over-expression of recombinant GPA1 so that the structure can also be determined experimentally.

ÖZET

Bitki G-proteinlerinin keşfedilmeleri ve tanımlanmaları memeli sistemlerdeki proteinlerle olan dizi benzerliklerine dayanmaktadır. Bitkiler üzerinde yapılan mutasyon çalışmaları bu proteinlerin işlevleri ile olası sinyal iletim mekanizmalarını ortaya çıkarmıştır (Fujisawa Y *et al.*, 2001). Bu tezde, *A. thaliana* G protein α alt biriminin (GPA1) detaylı dizi analizleri ile fare transdusin α alt birimiyle (PDB kod: 1gg2) karşılaştırmaları verilmektedir. GPA1 proteininin diğer bitki G protein α alt birimleriyle olan eşleştirmeleri yüksek oranda benzerlikler göstermektedir. Aynı analizler *A. thaliana* β ve γ alt birimleri içinde gerçekleştirilmiştir.

Dizi analizlerinin sonuçları ikincil yapı belirleme çalışmalarında ve GPA1 proteinin 3 boyutlu modelinin oluşturulmasında kullanılmıştır. İkincil yapı belirleme çalışmaları β ve γ alt birimleri içinde gerçekleştirilmiştir.

Model hesaplamaları dizi eşleştirmelerine, ikincil yapı belirleme çalışmalarına ve yapısı bilinen fare transdusin α alt birimine dayanmakta olup Insıght II programının içinde yer alan MODELLER modülü kullanılarak yapılmıştır. ERRAT kullanılarak yapıları kontrol edilen modellerin en iyisinin %76 oranında doğru olduğu belirlenmiştir. Bu modelin doğruluğu yapı içerisindeki değişken bölgelerin optimize edilmesi ile arttırılabilir. GPA1 proteinin yapısının deneysel olarak belirlenebilmesi için rekombinan protein ekspresyon çalışmaları sürmektedir.

To my family

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Assoc. Prof. Zehra Sayers, for her guidance and advice throughout this study. Without her support, I would not have been able to finish this work. It was a great relief during our hard days at Sabancı University, to have a supervisor and friend like her.

I would like to thank Assist. Prof. Ugur Sezerman, for his contributions and guidance to this thesis in the last month. It would have been harder to finish this work, without his knowledge and criticism.

I would like to express my special thanks to Suphan Bakkal, for her friendship, patience and hard work in the last two years. It would have been impossible to finish this work with another lab partner. I am also grateful to my friends Umit Ozturk, Kıvanc Bilecen, Ozgur Kutuk, Melis Tiryakioglu for their support, help and guidance through the past years.

Finally, I would like to thank faculty members and students at the Biological Sciences and Bioengineering Program, for making things a lot easier.

TABLE OF CONTENTS

ABBREVIATIONS

AC: Adenylyl cyclase

C-terminus: Carboxyl terminus

ERRAT: 3D structure evaluation server

Gα: G-protein alpha subunit

GAIP: G_{α} interacting protein

Gβ: G-protein beta subunit

GDP: Guanosine di-phosphate

Gγ: G-protein gamma subunit

GPA1: G_{α} protein from *A. thaliana*

GPA1: G_{α} gene from *A. thaliana*

GPCR: G-protein coupled receptor

GTP: Guanosine tri-phosphate

GTPase: enzyme converting GTP into GDP

Gα-GDP: Gα bound to GDP, in its inactive state

Gα-GTP: Gα bound to GTP, in its active state

META: Protein prediction server

N-terminus: Amino terminus

PSIpred: Secondary structure prediction server

PSSP: Secondary structure prediction server

RGS: Regulators of G-protein signaling

SAM-T99: Secondary structure prediction server

WD40 repeat: Tryptophane-Aspartate repeat consisting of 40 residues

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

G-proteins constitute a large family of regulatory GTP hydrolyzing proteins including Ras proteins, elongation factors and heterotrimeric G-proteins. All proteins of this family share a core structural organization. The GTP-hydrolysis mechanism is a feature of the core structure.

Heterotrimeric G-proteins consist of α, β and γ subunits. These subunits are complexed to each other at the interior part of the cell membrane, until being activated by a signal. The signal is transmitted from the outside by a receptor known as G-protein coupled receptor (GPCR). Upon activation, α subunit dissociates from βγ complex, enabling both Gα and Gβ/Gγ to bind to their effector systems and transmit the signal further.

At the core of every G-protein, is a guanine nucleotide-binding domain. The five-polypeptide loops (G-1, G-2, G-3, G-4, and G-5), that form the guanine nucleotide-binding site, are the most highly conserved elements in the domain and define the G-protein superfamily. The G-1 box contacts the α- and β-phosphates of the guanine nucleotide. The G-3 loop links the subsites for binding of Mg^{+2} and the γ-phosphate of GTP. The guanine ring is recognized, by the G-4 and G-5 loops.

After the advances in protein and nucleic acid sequencing methods, sequence databases have grown rapidly. In contrast, determination of protein structure by NMR or X-Ray crystallography has proceeded much more slowly. Hence, there are many proteins where the three-dimensional structure is not known, while the sequences are available through the databases.

Predicting the overall fold of a protein solely from its sequence is a major problem in current computational and structural biology fields. To overcome this problem, homology-modeling methods, that are able to predict the 3D structure of a protein sequence, were developed.

2 OVERVIEW

2.1 Small GTP-Binding Proteins

Small GTP-binding proteins are monomeric G-proteins that exist in eukaryotes from yeast to human and constitute a superfamily consisting of more than 100 members (Bourne HR *et al.*, 1990). The members of this superfamily are structurally classified into at least five families: the Ras, Rho, Rab, Sar1/Arf, and Ran families. Ras proteins mainly regulate gene expression, while the Rho/Rac/Cdc42 proteins of the Rho family regulate both cytoskeletal reorganization and gene expression. The Rab and Sar1/Arf family members regulate intracellular vesicle trafficking; and the Ran family members regulate nucleocytoplasmic transport during the G1, S, and G2 phases of the cell cycle and microtubule organization during the M phase.

Like heterotrimeric G-proteins, small G-proteins are found in two forms, GDP-bound inactive and GTP-bound active. Activation by an upstream signal, coming from both Gα and Gβγ subunits, leads to dissociation of GDP from the protein, followed by GTP binding (Bourne HR *et al.*, 1990; Hur EM *et al.*, 2002). Once activated, they bind to their downstream effectors and function as biological timers, unlike heterotrimeric G-proteins that function as molecular switches (Matozaki T *et al.*, 2000). The GTP-bound form is converted by the action of intrinsic GTPase activity to the GDP-bound form, which then releases the bound downstream effectors. This step, being the rate-limiting step, is extremely slow and requires a regulator, namely the guanine nucleotide exchange protein (GEP) or guanine nucleotide exchange factor (GEF) or guanine nucleotide releasing factor (GNRF1).

2.2 Heterotrimeric G-Proteins

GTP-binding proteins have the ability to efficiently bind and subsequently hydrolyze guanine nucleotides. Among these proteins, the heterotrimeric G-proteins are the only family taking role in signal transduction pathways (Bischoff *et al.*, 1999). The heterotrimer is formed by three subunits: α (alpha), β (beta), and γ (gamma); and is located on the cytoplasmic side of the plasma membrane. G-protein trimers respond to signals generated by transmembrane receptors called as G-protein coupled receptors (GPCR) that recognize ligands as diverse as glycoprotein hormones, cathacolamines, and light, by activating or inhibiting intracellular effector molecules, such as adenylyl cyclases, phospholipase Cβs, and ion channels (Downes GB *et al.*, 1999). Upon activation, α subunit dissociates from Gβ/Gγ complex, enabling both Gα and Gβ/Gγ to bind to their effector systems and transmit the signal further. Heterotrimeric G-protein subunits have also been identified in plants. Though plant G proteins have mostly different activators, they probably use the same or similar second messenger cascades that have to be identified. Despite their differences, plant G proteins have the same mechanism of activation as their mammalian counterparts (Bischoff F *et al.*, 1999).

2.2.1 Subunit families

In mammalians, the family of heterotrimeric G-proteins includes 23 isoforms from 4 classes of α (Gi, Gs, Gq, and G12), 5 of β and 12 of γ, known so far. In recent years, 13 Gα, 10 Gβ and 2 Gγ subunits have been identified and cloned in plants (Table 2.1). When signaling, they function in essence as dimers because the signal is communicated either by the Gα subunit or by the Gβγ complex. These interact specifically with over a thousand of different receptors and more than a dozen effectors (Downes GB *et al*., 1999).

Table 2.1 Plant G protein designations *(Assmann SM, 2002).*

Gene	Species	Classification	Reference
GPA1	Arabidopsis	Gα	Ma et al., 1990
TGA1	Tomato	Gα	Ma et al., 1991
LjGPA1	Lotus	Gα	Poulsen et al., 1994
RGA1/D1	Rice	Gα	Ishikawa et al., 1995; Seo et al., 1995
SGA1	Soybean	Gα	Kim et al., 1995
SGA2	Soybean	Gα	Gotor et al., 1996
NtGPα1	Tobacco	Gα	Saalbach et al., 1999
NtGA2	Tobacco	Gα	Ando et al., 2000
LGPα1	Lupin	Gα	Kusnetsov and Oelmueller, 1996b
AfGα1	Wild oat	Gα	Jones et al., 1998
PGA1, PGA2	Pea	Gα	Marsh and Kaufman, 1999
SOGA1	Spinach	Gα	Perroud et al., 2000
NPGPA1	*Nicotiana plumbaginifolia*	Gα	Kaydamov et al., 2000
AGB1	Arabidopsis	Gβ	Weiss et al., 1994
ZGB1	Maize	Gβ	Weiss et al., 1994
TGB1	Tobacco	Gβ	Kusnetsov and Oelmueller, 1996a
RGB1	Rice	Gβ	Ishikawa et al., 1996
AfGβ1	Wild oat	Gβ	Jones et al., 1998
AfGβ2	Wild oat	Possible Gβ	Jones et al., 1998
NPGPB1	*Nicotiana plumbaginifolia*	Gβ	Kaydamov et al., 2000
AGG1	Arabidopsis	Gγ	Mason and Botella, 2000
AGG2	Arabidopsis	Gγ	Mason and Botella, 2001
GCR1	Arabidopsis	Potential heterotrimeric G protein receptor	Josefsson and Rask, 1997; Plakidou-Dymock et al., 1998
MLO	Barley	Potential heterotrimeric G protein receptor	Devoto et al., 1999
AtXLG1	Arabidopsis	Extra large GTP binding protein	Lee and Assmann, 1999
PsDRG	Pea	Developmentally regulated G protein	Devitt et al., 1999
AtDRG	Arabidopsis	Developmentally regulated G protein	Etheridge et al., 1999; Devitt et al., 1999
RDH3	Arabidopsis	(putative GTP-binding protein)	Schiefelbein and Somerville, 1990
ATGB1	Arabidopsis	GTP-binding protein	Biermann et al., 1996
fw2.2/ORFX	Tomato	(putative GTP-binding protein)	Frary et al., 2000

2.2.1.1 The α-subunit

Being the major signaling component, the Gα subunits in mammalians have been divided into four families based on sequence similarities, while Gβ and Gγ subunits have not been classified (Neves SR *et al.*, 2002). (Table 2.2)

Table 2.2 G-protein subfamilies and their activities *(Neves SR et al., 2002).*

G-protein subfamily	Family members	Properties	Activities
$G\alpha_s$	$G\alpha_{s(1-4)}$, $G\alpha_{olf}$	Cholera toxin activates	Stimulate adenylyl cyclase Open calcium channels
$G\alpha_{i/o}$	$G\alpha_{i(1-3)}$	Pertussis toxin inhibits	Inhibit adenylyl cyclase Open potassium channels
	$G\alpha_{o(1,2)}$	Pertussis toxin inhibits	Close calcium channels
	$G\alpha_z$		Inhibit adenylyl cyclase
	$G\alpha_{t(1,2)}$	Pertussis, cholera toxin sensitive	Activate cGMP phosphodiesterase
	$G\alpha_{gust}$		
$G\alpha_q$	$G\alpha_q$, $G\alpha_{11}$ $G\alpha_{14}$, $G\alpha_{15}$, $G\alpha_{16}$		Stimulate PLCβ
$G\alpha_{12}$	$G\alpha_{12}$, $G\alpha_{13}$		?
$G\alpha_h$	$G\alpha_h$		Stimulate PLCβ
$G\beta\gamma$	$G\beta_{1-5}$, $G\gamma_{1-7}$		Stimulate PLCβ Activate or inhibit adenylyl cyclases

The Gs and Gq families have very well defined effector pathways, the adenylyl cyclase and phospholipase C-β pathways, respectively. The Gi and Go families are less structured, and here the signal flows through both the Gα and Gβγ complexes. Although signaling with the G12 and G13 pathways have been extensively studied, with similar downstream components being identified, it is not clear whether they always regulate similar biological functions. (Figure 2.1)

G-protein α subunits range from 350 to 395 amino acids in length. Plant α subunits have at least 70% homology, while mammalian α subunits show at least 40% homology and the levels of homology between plant and mammalian Gα subunits were as high as 38-40% (Lochrie MA, 1988).

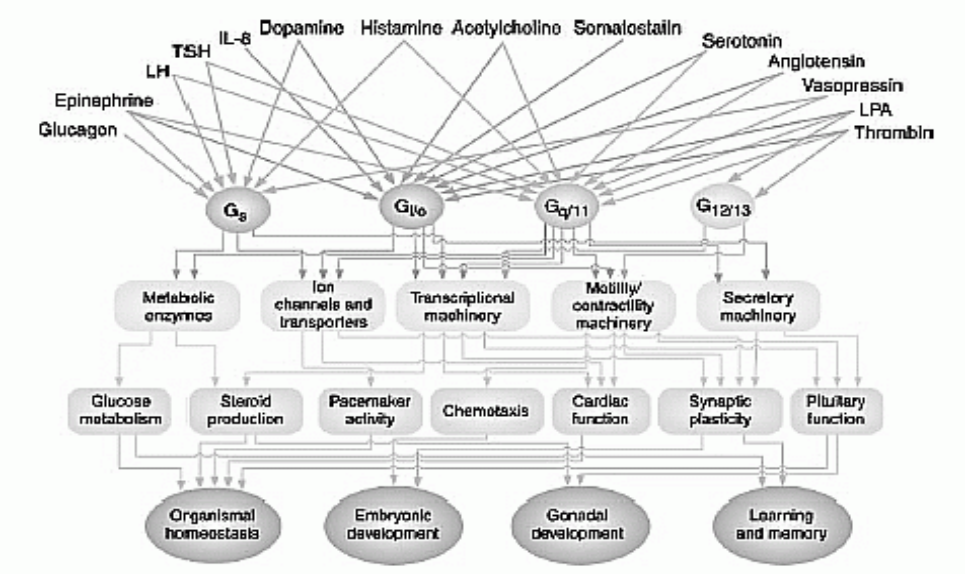


Figure 2.1 G-protein pathways *(Neves SR et al., 2002).*

2.3 G-protein Coupled Receptors

Precise control of physiological phenomena is performed by various kinds of receptor-mediated signaling. The vast number of receptors belongs to the GPCR family, making them one of the largest receptor families found in nature. Recent estimates are that about 1% of the mammalian genome code for these types of receptors, and thousands of GPCRs are predicted to exist. (Hur *et al.*, 2002) Plant GPCRs have also been identified and cloned (Josefsson LG, 1997).

GPCRs have a common structure formed of seven transmembrane helices, connected by intracellular loops that form the G-protein-binding domain. Ligand binding to the receptor causes conformational changes, resulting in the activation of GPCR-interacting proteins, but the exact mechanism of this process is not completely understood yet (Hur *et al.*, 2002; Brady *et al.*, 2002). Models describing the interaction of GPCRs with their protein targets were generally based on the assumption that the receptors exist as monomers and couple to G-proteins in a 1:1 stoichiometric manner. However, recent studies have shown both the existence of multi-domain scaffolding

proteins and chaperone molecules interacting with GPCRs, and homo- and hetero-oligomerization of the receptors themselves, generating numerous possibilities for the role and function of GPCRs than has been estimated (Brady *et al.*, 2002).

2.4 Regulators of G-Protein Signaling

Regulators of G-protein signaling (RGS) proteins act as negative regulators of G-protein dependent signaling, as they enhance GTP-hydrolysis by Gα subunits to turn off signaling (Dohlman HG, 1997; Wilkie TM, 2000). These proteins form a highly diverse family, having unique tissue distributions, and strong regulation by signal transduction events.

In addition to their GTPase-accelerating activity, RGS proteins also: directly antagonize Gα effectors (Hepler *et al.*, 1997), enhance receptor-G-protein coupling (Zhong H, 2001), are G13 effectors (Kikuchi A, 1999), may participate in desensitization or tolerance to opioids (Potenza MN *et al.*, 1999).

A member of RGS family that is well defined is the Gα-interacting protein (GAIP). These proteins are able to regulate heterotrimeric G-protein activity by specific binding to Gα subunits (Wylie F *et al.*, 1999).

2.5 Mechanism of Activation and Action

In the inactive state, G-proteins form membrane-bound αβγ heterotrimers, with GDP tightly bound to the α-subunit. Receptor activation by ligand binding is thought to cause changes in the relative orientations of transmembrane helices 3 and 6, though high resolution structure data has not been reported for GPCRs, yet (Hamm HE, 1998). These changes then affect the conformation of G-protein interacting intracellular loops

of the receptor and thus may uncover previously blocked G-protein-binding sites (Farrens D *et al.*, 1996).

When an activated receptor interacts with the Gα, the exchange of GDP for GTP is catalyzed (Perroud *et al.*, 2000). GTP binding leads to rapid dissociation of Gα-GTP from the Gβγ complex, due to the unstable nature of GTP-bound form of the heterotrimer. Activation of downstream effectors occur both by Gα-GTP and Gβγ subunits. G-protein deactivation is the rate-limiting step for switching off the cellular response and occurs when the α-subunit hydrolyzes GTP to GDP (Zimmermann H, 1993; Hamm HE, 1998). This GTPase activity is activated by the binding of the α subunit to its effector and reforms the original complex with the βγ subunits.

Certain oncogenes have been shown to code for G-proteins that contain a defect in their intrinsic GTPase activity. The oncogene product Ras is a good example. Such proteins, once bound to GTP, are unable to return to the inactive state, causing the cell to continue to receive the 'on' signal, even in the absence of receptor binding to a ligand. On the other hand, the toxin released from the bacterium *Vibrio cholerae*, known as cholera toxin, inactivates the α subunit of G-proteins. Being an oligomeric enzyme, the toxin catalyzes the transfer of ADP-ribose to a specific arginine residue of the α subunit, destroying its GTPase activity. By this way, the G-protein becomes irreversibly activated and the effector systems become either activated or deactivated permanently.

Heterotrimeric G-proteins transduce ligand-induced signals into intracellular responses, which underlie physiological responses of tissues and organs. G-proteins play important roles in determining the specificity of the cellular responses to signals (Hamm HE, 1998). However, because of the diversity of G-protein subunits and downstream effector molecules, the pattern of responses of a particular cell or tissue to stimulation by a ligand is quite complex and only a general scheme can be given to explain this process.

Once activated, both α and βγ subunits interact with their effector systems. Different types of Gα's interact with various effector systems, including adenylyl cyclase, cGMP phoshodiesterase, and phospholipase C-β. Moreover, α subunits have

been shown to interact with other proteins, like RGS and GAIP. Interestingly, βγ subunits use some of the effector molecules that the α subunit uses, namely adenylyl cyclases and phospholipases. (Figure 2.2)

However, Gβγ subunits mediate signal transduction by interaction with many other proteins including GPCRs, GTPases, K^+ channels, voltage-sensitive Ca^{+2} channels, PI3 Kinase, and molecules within the MAPK pathway (Hur EM *et al.*, 2002).

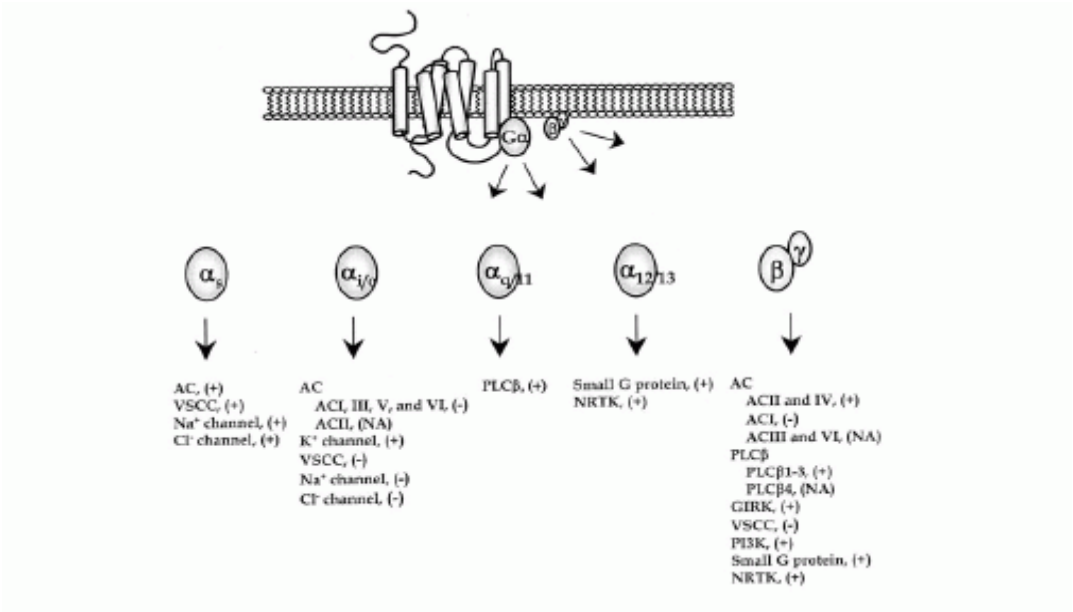


Figure 2.2 Effector molecules regulated by different Gα and Gβγ subunits *(Hur EM et al., 2002).*

2.5.1 Adenylyl Cyclase

In mammals, adenylyl cyclase is a single polypeptide that resides in the plasma membrane of cells. Its role is to catalyze production of cAMP from ATP. The 3`-OH ribose group of the ATP attacks the α-phosphoryl group resulting in cyclisation of the molecule. The energy released from this reaction helps to drive the cyclisation reaction (Tang WJ, 1992).

Like many of the proteins in signal transduction, there are various isoforms of adenylyl cyclase. Although the mammalian ones are integral in the plasma membrane, other forms have been found to be peripheral plasma-membrane proteins, for example in the yeast *S. cerevisiae* or in E. coli. At least eight isoforms have been identified in mammals (type I-VIII), but all share similar structural topology. They contain two clusters of six transmembrane-spanning highly hydrophobic domains that separate two catalytic domains on the cytoplasmic side of the membrane.

When a hormone binds to the relevant receptor on the cell surface, activation of adenylyl cyclase is an indirect process, requiring the use of G-proteins. The action of a G-protein on the activity of adenylyl cyclase is either stimulatory or inhibitory, depending on the G-protein involved. Stimulatory effect results in increased cAMP levels, while the inhibitory effect causes the opposite (Mattera R *et al.*, 1989; Cooper DMF *et al.*, 1995). However, G-proteins are not the only means of controlling the activity of adenylyl cyclases. Ca^{+2}/Calmodulin can also activate some of the isoforms of the enzyme (Cooper DMF *et al.*, 1995).

Although the active part of trimeric G-proteins has been thought to be the α subunit, the βγ complex has also been shown to be important (Ahmed AH *et al.*, 1997). Moreover, it was shown that different isoforms of adenylyl cyclase were controlled differently by the subunits of the G-protein heterotrimer. While some adenylyl cyclases were stimulated by the βγ complex, others were inhibited or even not affected (Iñiguz-Lluhi J *et al.*, 1993).

2.5.2 Phospholipase C-β

One of the key events in signal transduction in cells takes place on the membrane and involves the breakdown of some of the lipids of the membrane. Phospholipase C-β (PLC) enzymes belong to a large family of inositol-lipid specific phospholipase C's with well-defined functions. They hydrolyze several inositol lipids including phosphatidylinositol-4, 5-biphosphate (PIP_2), phosphatidylinositol (PI), and phosphatidylinositol 4-phosphate (PI4-P). Hydrolysis of PI and PI4-P releases the

inositol phosphates such as inositol-1, 4, 5- triphosphate (IP_3) and diacylglycerol (DAG) (Divecha N *et al.*, 1995).

IP_3 is responsible for the release of Ca^{+2} ions from intracellular stores, which leads to the activation of the Ca^{+2} signaling pathways, including the activation of calmodulin and its associated effectors. DAG, on the other hand, leads to the activation of protein kinase C (PKC) and the associated phosphorylation of a host of proteins along with modulation of their activity.

Activation of PLC is brought about in different ways depending on the isoforms, but one of the mechanisms for turning on PLC is through the interaction with components of the trimeric G-proteins. Both α and βγ subunits of heterotrimeric G-proteins can bind and affect PLC-β activity. Furthermore, it has been shown that the binding sites on PLC-β for α and βγ subunits are different, indicating that they can interact simultaneously (Smrcka AV *et al.*, 1993).

PLC's have also been reported to control the signal transduction cascade in which they are involved. This is achieved through their GTPase-activation activity, exerted on α subunits of G-proteins. Upon activation by G-proteins, PLC-β activates its downstream effectors and binds to the α subunit to improve GTP-hydrolysis (Cook B *et al.*, 2000; Montell C., 2000).

2.5.3 Ion channels

Ion channels are encoded by several hundred genes in mammalians, differing widely in molecular structure, selectivity to ions and mechanisms of operation. Despite the structural diversity, these proteins share a general structural motif: a pore formed by and enclosed within the transmembrane segments of the channel protein, through which ions traverse the plasma membrane.

Ion channels are end targets in a large number of regulatory pathways that are initiated by G-protein coupled pathways. Both α and βγ subunits can regulate ion channels directly, via physical interactions between G-protein subunits and the channel protein, or indirectly, via second messengers and protein kinases (Wickman K *et al.*, 1995).

Direct modulation by G-proteins has been proposed mainly for two families of ion channels, voltage-dependent Ca^{+2} channels and G-protein-activated inwardly rectifying K^{+} channels (GIRKs or Kir3) (Dascal N, 2001).

2.6 G-protein Functions In Plants

Heterotrimeric G-proteins have distinct and important roles in various organisms, including bacteria, fungi, plants and mammalians. In mammalians, heterotrimeric G-proteins have been reported to play a role in taste perception (McLaughlin SK *et al.*, 1992), visual transduction (Arshavsky VY, 2002), several physiological disorders including Alzheimer's disease (Cowburn RF *et al.*,2001), hormonal signaling and immune system (Kehrl JH, 1998; Lania A *et al.*, 2001) and development (Malbon CC, 1997).

Plant G-proteins are thought to take role in various signaling processes, including plant hormone signaling, light signaling, pathogen signaling, development, seed germination and growth, regulation of biosynthetic pathways as well as regulation of ion (K^{+} and Ca^{+2}) channels and opening of stomatal guard cells (Fujisawa Y *et al.*, 2001; Millner PA, 2001).

Although studies using mutating agents activated or inhibited G-protein pathways, the exact mechanism of G-protein involvement within the plant hormone signaling is still not clear. In plants, G-proteins have been reported to take part in signal transduction events of the hormones auxin, gibberellin, and abscisic acid (ABA) (Millner PA, 2001).

Auxin is mainly produced in apical meristems, buds, young leaves, and other active young parts of plants. Besides stimulating the enlargement of cells by increasing the plasticity of cell walls, auxins have many other effects, including triggering the production of different growth regulators, causing the Golgi bodies to increase rates of secretion, playing a role in controlling some phases of respiration, and influencing numerous developmental aspects of growth (Stern KR, 1997).

Gibberellin, being a plant hormone and a growth regulator, controls diverse developmental processes as seed germination, stem elongation, leaf expansion, trichome development, flower and fruit development (Davies PJ, 1995).

ABA is synthesized in plastids, apparently from carotenoid pigments. It is found in many plant materials but is particularly common in fleshy fruits, where it evidently prevents seeds from germinating while they are still on the plant. As the stimulatory effects of other hormones are inhibited by ABA, cell growth is also inhibited (Stern KR, 1997).

In higher plants, guard cell ion-channel regulation controls stomatal apertures. Stomatal opening relies on increases in K^+, Cl^-, malate^{2-}, and sucrose in the guard cell symplast to drive water influx and cell swelling. These processes result in an outbowing of the guard cell pair and an increase in pore aperture. During stomatal opening, K^+ uptake is mediated by inwardly rectifying K^+ channels. During inhibition of stomatal opening by ABA, these channels are inhibited. In guard cells, ABA activates phospholipases C and D and can elevate cytosolic calcium levels through IP_3 or other pathways (MacRobbie EAC, 2000; Leckie SP *et al.*, 1998). Cytosolic Ca^{+2} elevation, in turn, inhibits inwardly rectifying K^+ channels and activates slow anion channels that mediate Cl^- and malate^{2-} efflux (Wang X-Q *et al.*, 1998). It has been shown that regulation of ion channels in stomatal guard cells and ABA, at least in part, involves G-proteins in plants (Wang X-Q *et al.*, 2001).

2.7 Structural Features of G-proteins

At the core of every G-protein, is a guanine nucleotide-binding domain. The five-polypeptide loops (G-1, G-2, G-3, G-4, and G-5), that form the guanine nucleotide-binding site are the most highly conserved elements in the domain, and define the G-protein superfamily. (Table 2.3)

The diphosphate binding loop (P-loop or G-1 box), having the consensus sequence GXXXXGK(S/T) connects the β1 strand to the α1 helix and contacts the α- and β-phosphates of the guanine nucleotide. The G-3 loop, with the consensus sequence DXXG, at the N-terminus of the α2 helix links the subsites for binding of Mg^{+2} and the γ-phosphate of GTP. The guanine ring is recognized, in part, by the G-4 loop that links the β5 strand and the α4 helix. The connection between the α1 helix and the β2 strand (G-2) contains a conserved threonine residue involved in Mg^{+2} coordination. G-5, located between β6 and α5, forms the guanine base recognition site (Sprang SR, 1997).

2.7.1 Heterotrimeric G-protein α subunits

The general structure of Gα, in mammalian systems, is formed of two domains: a GTPase domain and a helical domain. The GTPase domain contains the guanine nucleotide binding pocket, the Mg^{2+}-binding domain, the guanine ring binding motif, the threonine and glycine residues needed for GTP hydrolysis, an N-terminal lipid modification site, and sites for binding receptors, effector molecules, and the βγ subunit. The helical domain contains the arginine residue needed for GTP hydrolysis, and is thought to slow down GTP hydrolysis but its exact function is not clear yet. Plant Gα subunits are thought to have almost the same structure, as there are still no structural studies conducted in plants.

Effector interaction involves three regions within the GTPase domain. The first of these regions, called Switch I, is a loop connecting helix α4 to strand β6. The second region, namely Switch II, corresponds to the loop preceding the α2 helix, and the helix

itself. Switch III, the last region, corresponds to the loop that connects helix α3 to strand β5. (Figure 2.3)

The base of the β propeller is positioned directly over the β_2-β_3-α_2 cluster of the α subunit. All but three of the DA (β blades 6 and 7) and BC (blade 6) loops contribute to the α contact surface. Almost the entire length of the switch II region of the α subunit, which corresponds to the β_3-α_2 loop and the α_2 helix, is buried in the contact with β subunit. The hydrophobic core of the contact is organized around Trp-99 of β and Trp-211 of the α subunit. This contact also includes the β subunit residues Tyr-59, Met-101, Leu-117, Tyr-145, and Met-188 and the α subunit residues Ile-184, Phe-199, Cys-214, Phe-215, and Lys-210. Trp-99 of the β subunit protrudes into a hydrophobic pocket in the surface of the α subunit (Wall MA *et al.*, 1995). Mutational studies involving the Trp-99 of the β subunit resulted in disruption of the interaction between α and βγ subunits (Whiteway M *et al.*, 1994).

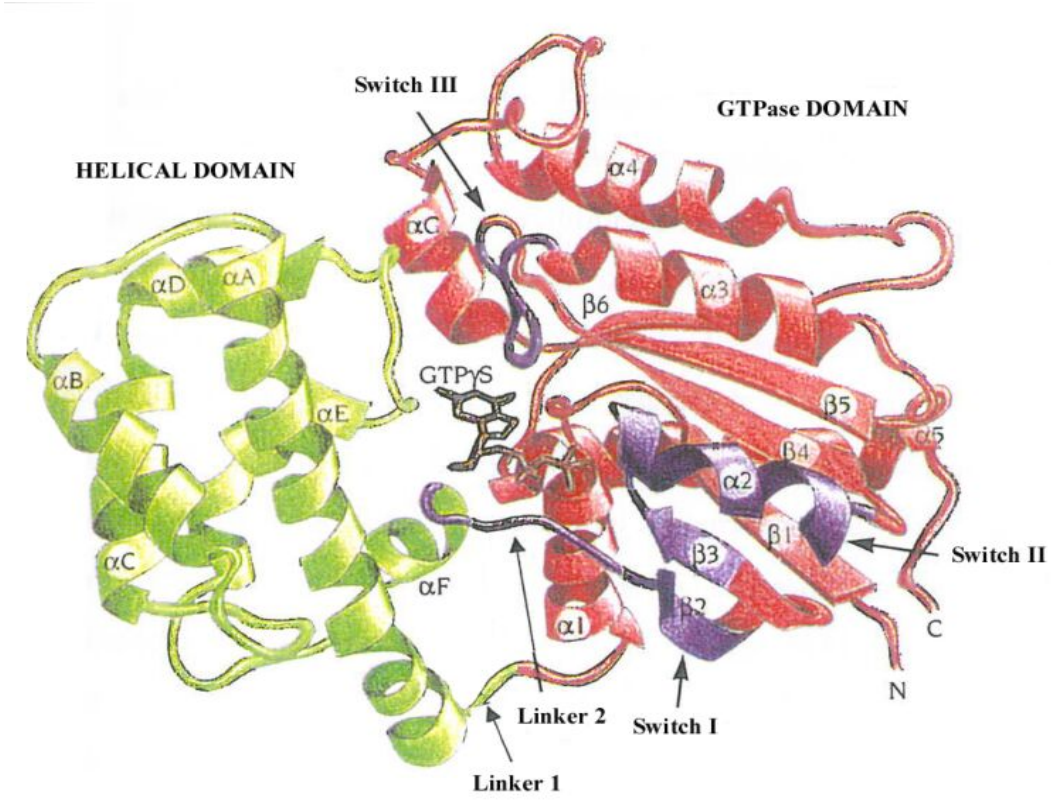


Figure 2.3 Overall structure of G-protein α subunit.

(Rens-Domiano S et al., 1995)

The GTP-hydrolysis mechanism involves a conformational change of the α subunit, thereby enabling the formation of the heterotrimer. Both the switch II and the amino-terminal parts of the α subunit are dynamic components of the interaction. Upon GTP-hydrolysis, the switch II helix rotates ~120^{o}, exposing the hydrophobic residues Phe-199, Trp-211, and others, to interact with complementary nonpolar pockets in the β subunit. The same rotation also creates two ionic interactions between the α and β subunit: Glu-216 on α forms an ion pair with Lys-57 on β and Lys-210 is inserted into a negatively charged pocket formed by Asp-288 and Asp-246 on adjacent loops of the β subunit (Sondek J *et al.*, 1994).

Formation of the α/β interface likewise destabilizes the conformational state of switch II that is required for GTP binding. The amino-terminus of the helix, of the switch II region, forms the binding site for the γ phosphate of GTP and positions the catalytic residue Gln-204 for a role in transition state stabilization. The conformational changes within switch II region are coordinated with a complementary shift of the switch I peptide that ultimately traps GDP in the catalytic site of α subunit (Sondek J *et al.*, 1996).

In the GTP-bound state, Thr181 of switch I contributes an oxygen ligand to the Mg^{+2}. This structural constraint is removed with the release of Mg^{+2} upon GTP-hydrolysis. Switch I region is removed from the catalytic site, to avoid any interactions with Gly-203 caused by rotation of the switch II helix (Coleman DE *et al.*, 1994). This results in the repositioning of Glu-186 and Ile-184 for optimal interaction with the β subunit, and promotes hydrogen bonding as well as ion pair formation between Arg-178 and Glu-43. Formation of the ion pair prevents the diffusion of the nucleotide from the catalytic site. Thus, Arg-178 is crucial for catalysis, and binding of βγ and GDP to the α subunit (Higashijima T *et al.*, 1987). In the heterotrimer, the carboxyl-terminus of the α subunit is exposed for possible interactions with the cytoplasmic domains of their receptors (Conklin BR *et al.*, 1993).

2.7.2 Heterotrimeric G-protein β subunits

The core of the β subunit folds into a β-propeller domain that is composed of seven repeats, termed as WD40 repeats, comprising a four-stranded antiparallel sheet that is called the β blade. WD40 repeats are approximately 40 amino acids in size, with several additional conserved amino acids, including a Trp-Asp dipeptide. The WD40 repeat corresponds to the outer strand of one β blade and the first three inner strands of the next (Wall MA *et al.*, 1995).

Each of the seven β blades is organized around a narrow central channel, with the β strands roughly parallel with the channel axis. The α subunit is positioned at the narrow end of the channel, forming two contact surfaces with the β subunit. The first of these includes the cluster containing the β2 and β3 strands and the α2 helix in the α subunit. This segment contains the switch II region that undergoes conformational rearrangement upon hydrolysis of GTP. The second contact region is an extended interface between the helical amino terminus of α subunit and the side of the β propeller (Lambright *et al.*, 1994).

To form the β subunit, the seven four-stranded sheets (β blades) are nested face to face around the barrel axis. The first strand (A) of the β blade lines the inner channel of the structure, and the rest advances outward so that the fourth strand (D) forms the outer edge. The AB loops and the CD loops are located at the end, where the γ subunit is bound, while the BC and long DA loops are located on the opposite end that faces the α switch II region. The amino terminal of the β subunit forms an extended polypeptide chain that tightens the top of the barrel. This part of the chain forms a shallow groove on the surface of the barrel that binds the γ subunit. (Figure 2.4)

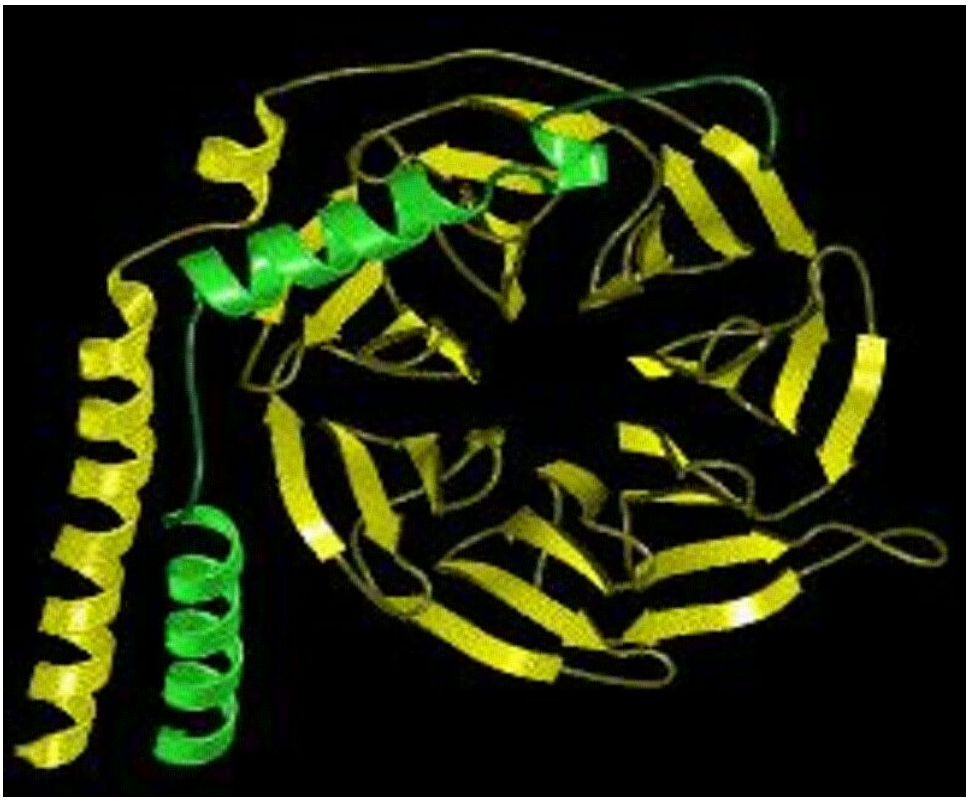


Figure 2.4 Overall structure of G-protein β and γ subunits. β subunit is shown in yellow, γ subunit in green. *(Tooze J and Branden C, 1999)*

2.7.3 Heterotrimeric G-protein γ subunits

The γ subunit of the heterotrimeric G-protein is composed of two helical segments joined by a loop and has essentially no tertiary structure. The direction of the γ chain is parallel to that of the amino-terminus of the β subunit. A part of the amino-terminal of the γ chain is α-helical and crosses the amino-terminal helix of the β subunit at an angle of 15^{o}, suggestive of a parallel coiled-coil interaction (Garritsen A *et al.*, 1993).

Like the β subunit, the γ subunit projects into solvent towards the α subunit, where the terminal residues are disordered. The second helix in the γ subunit runs side by side with the second helix of the β subunit, covering the fifth and sixth blades of the β subunit. The carboxyl-terminal loop of γ subunit is buried in a hydrophobic pocket on the surface of the β subunit (Katz I *et al.*, 1995).

Table 2.3 Conserved nucleotide-binding motifs in selected G-proteins *(modified from Sprang SR, 1997).*

Organism	Protein	Residue	G-1	Residue	G-2	Residue	G-3	Residue	G-4	Residue	G-5
Human	Ras	10	**G**AGGV**GKS**	32	YPD**T**IED	55	IL**D**TA**G**QE	114	VG**NKCD**	142	YIET**SAK**
Human	Rap-1A	10	**G**SGGV**GKS**	32	YDP**T**IED	55	IL**D**TA**G**TE	114	VG**NKCD**	140	FLGS**SAK**
Human	ARF-1A	24	**G**LGAA**GKT**	45	TIP**T**IGF	65	VW**D**VG**G**QD	124	FA**NKQD**	155	IQAT**CAT**
E. coli	EF-Tu	18	**G**HVDH**GKT**	58	RGI**T**INT	79	HV**D**GP**G**HA	133	FL**NKCD**	169	IVRG**SAL**
T. thermophilus	EF-G	19	**A**HIDA**GKT**	61	RGI**T**ITA	81	II**D**TP**G**HV	135	FA**NK**M**D**	258	VFLG**SAL**
Bovine	$G_{s\alpha}$	47	**G**AGES**GKS**	201	RVL**T**SGI	221	MF**D**VG**G**QR	290	FL**NKQD**	361	PHFT**CAV**
Bovine	$G_{i\alpha}$	40	**G**AGES**GKS**	178	RVK**T**TGI	198	LF**D**VG**G**QR	267	FL**NKK**D	321	THFT**CAT**
Bovine	$G_{t\alpha}$	36	**G**AGES**GKS**	174	RVKTTGI	194	MF**D**VG**G**QR	263	FL**NKK**D	317	SHMT**CAT**
Bovine	$G_{o\alpha}$	40	**G**AGES**GKS**	179	RVK**T**TGI	199	MF**D**VG**G**QR	268	FL**NKK**D	321	CHMT**CAT**
Human	$G_{z\alpha}$	40	**G**TSNS**GKS**	179	RDMTTGI	199	MF**D**VG**G**QR	268	FL**NKK**D	322	SHFT**CAT**
Mouse	$G_{q\alpha}$	40	**G**TGES**GKS**	177	RVP**T**TGI	197	MF**D**VG**G**QR	266	FL**NKK**D	320	SHFT**CAT**
Tomato	Tga1	45	**G**AGDS**GKS**TI	191	RIR**T**TGV	217	LF**D**VG**G**QR	285	LFL**NKF**D	251	IYRT**TAL**
Potato	Stgpa1	45	**G**AGDS**GKS**TI	191	RIR**T**TGV	217	LF**D**VG**G**QR	285	LFL**NKF**D	251	IYRT**TAL**
N. plumbaginifolia	Npga	53	**G**AGDS**GKS**TI	199	RIR**T**TGV	225	LF**D**VG**G**QR	273	LFL**NKF**D	240	IYRT**TAL**
Spinach	SOGA1	44	**G**AGES**GKS**TI	190	RVR**T**TGV	216	LF**D**VG**G**QR	284	LFL**NKF**D	250	IYQA**TAF**
Arabidopsis	GPA1	44	**G**AGES**GKS**TI	190	RVR**T**TGV	216	LF**D**VG**G**QR	284	LFL**NKF**D	250	IYRT**TAL**

2.8 Homology Modeling

After advances in protein and nucleic acid sequencing methods, sequence databases such as the protein information resource (http://www-nbrf.georgetown.edu/), SwissProt and TrEMBL (http://www.expasy.ch/) have grown rapidly. In contrast, experimental determination of protein structure by NMR or X-Ray crystallography has proceeded much more slowly and there are many proteins where the three-dimensional structure is not known, but the sequences are available through the databases.

One of the major problems currently faced in computational and structural biology is to be able to predict the overall fold of a protein correctly, purely from its sequence. This is known as the protein-folding problem (Osguthorpe DJ, 2000). Homology modeling methods may predict the 3D structure of a protein sequence by using information derived from a homologous protein of known structure (Sanchez R, 1997).

In order to construct a homology model for a query protein sequence, the query must first be aligned with one or more homologous reference proteins of known structure. If the sequence identity between the two proteins within the alignment falls to 30% or below, the alignment process becomes increasingly unreliable (Venclovas C *et al.*, 1999) and results in incorrectly folded regions in the predicted structure.

There are two methods of homology modeling, which can be used: fragment-based homology modeling and restraint-based homology modeling. Fragment-based homology modeling procedures use the alignment between the query sequence and the known protein(s) to identify a number of structurally conserved regions (SCR). These regions have no insertions or deletions and tend to have well defined secondary structures, like helices or strands. The level of sequence conservations is also the highest at these sections of the protein. Regions of the protein sequences in between the structurally conserved regions are usually denoted as variable regions (Forster MJ, 2002).

As the SCRs provide a consistent framework between the known and unknown structures, the coordinates of the protein backbone in the query protein can be copied from those in one of the known proteins. Variable regions in an alignment, on the other hand, are most often protein loop regions where mutations, insertions, and deletions are common and these make accurate modeling hard. Generally, loops are modeled by searching a structural database for regions of a suitable length and geometry at the interface with the SCRs, along with the requirement that they do not have any interference on the rest of the model structure (Forster MJ, 2002). Several commercially available homology-modeling programs utilize this approach for model construction. These include the COMPOSER program (Blundell TL *et al.*, 1987, 1988; Topham CM *et al.*, 1990) which is incorporated into the SYBYL program suite (Tripos Inc., St. Louis, *http://www.tripos.com*) and HOMOLOGY, which is a component of InsightII (MSI, San Diego, *http://www.accelrys.com*).

Unlike fragment-based methods, the restraint-based homology modeling methods do not break the model building process into two distinct phases i.e. building conserved regions then finding variable loop regions. Instead, the alignment is used to derive geometrical restraints, such as limits on distances between pairs of C_α atoms, ranges of backbone, and side chain dihedral angles, etc. These restraints can then be combined together to find an overall scoring function that defines how well the model structure matches the set of geometric criteria. A structure generation procedure, of which multiple types have been reported, is then used to create model structures that best satisfy the restraints (Forster MJ, 2002). Restraint-based molecular dynamics procedures for structure generation have been used in the MODELLER program that is available for academic authors through the author's web site (*http://guitar.rockefeller.edu/modeller/modeller.html*) or commercially as a component of InsightII.

3 AIM OF THE STUDY

In this M.Sc. work, detailed sequence alignment analyses of *A. thaliana* G-protein α subunit (GPA1) with its mammalian counterpart rat transducin α subunit (PDB entry 1gg2 and 1fqk) were carried out. Alignment of GPA1 with other known Gα sequences, from plants, reveal a high degree of homology. Same analyses were extended to *A. thaliana* β and γ subunits.

Results of sequence analyses were used as a basis for secondary structure prediction and for modeling 3D structure of GPA1. Secondary structure prediction was carried out for β and γ subunits as well.

Furthermore, a model structure for GPA1 was proposed, based on the sequence alignment, secondary structure prediction and known structure of rat transducin α subunit.

4 MATERIALS

4.1 Chemicals

All chemicals and growth mediums were purchased from SIGMA (USA), Fluka (Switzerland), Merck (Germany), and Riedel de Häen (Germany).

4.2 Primers and Vectors

Primers for *GPA1*, without restriction enzyme cutting sites, designed according to the sequence reported by Ma H (Ma H *et al.*, 1990) were purchased from Integrated DNA Technologies Inc., USA.

G1-primer (forward): 5' – ATG GGC TTA CTC TGC AGT – 3'

G3-primer (reverse): 5' – TCA TAA AAG GCC AGC CTC – 3'

pGEM-T Easy (Promega) is a vector facilitating cloning of the nucleic acid fragment with its 3' Adenine tail.

4.3 Enzymes

T4 DNA ligase (Promega) – 3u/µl

Taq DNA polymerase (Promega) – 5u/µl

EcoRI Restriction enzyme (Promega) – 12u/µl

4.4 Commercial Kits

QIAGEN, Qiaquick® Gel Extraction Kit (250)

QIAGEN, Qiaquick® PCR Purification Kit (50)

QIAGEN, Qiaquick® PCR Purification Kit (250)

QIAGEN, Qiaprep® Spin Miniprep Kit (250)

QIAGEN, QIAGEN® Plasmid Midi Kit (100)

Promega, PCR Core System II

4.5 Buffers and Solutions

Luria-Bertani Growth Medium:

1% w/v tryptone

0.5% w/v yeast extract

1% w/v NaCl, pH: 7.2

4.6 Equipment

A complete listing of equipment used in this study can be found in Appendix B.

5 METHODS

5.1 Culture Growth

Basic culture growth, competent cell preparation, cloning, plasmid isolation, and restriction enzyme analysis procedures flowed were used according to Sambrook *et al.* (Sambrook J and Russell DW, 2001) throughout this study, *E. coli* JM109 cells were kindly provided by Prof. Beki Kan (Dept. of Biophysics, Marmara University, Istanbul, TÜRKİYE). The cells were grown overnight in Luria-Bertani (LB) medium, prior to any application.

5.2 Polymerase Chain Reaction (PCR)

The vector pCIT767 containing *GPA1* sequence was digested overnight at 37^{o} C using the restriction enzyme EcoRI. Three units enzyme/µg DNA was used for the digestion mix. PCR was carried out using PCR Core System II, *GPA1* being the template together with the designed primers (G1 and G3). The following program was used in the thermocycler (Eppendorf thermocycler):

Template: 3 µl (final: < 0.5 µg/50 µl)

Primer (G1 and G3): 1 µM final conc.

- 94^{o} C, 1 min.

- 94° C, 1 min.
- 53° C, 1 min.
- 72° C, 1 min.

} 30 cycles

- 72° C, 1 min.
- 22° C, hold.

5.3 Preparation of pGEM-T Easy + *GPA1* Construct

Ligation reaction was prepared using purified *GPA1* fragment as the insert, and cut pGEM-T Easy as the vector. The reaction mix was prepared according to the manufacturer's instructions with three µl insert. Besides the desired ligation reaction, controls with no insert DNA (background control) and with control insert DNA (positive control) were also prepared. The ligation mix was used to transform competent *E. coli* JM109 cells.

5.4 Plasmid Isolation and Verification of *GPA1* Sequence

GPA1 was kindly provided by Dr. Hong Ma (Dept. of Biology, Pennsylvania State University, Pennsylvania, USA) in a pCIT767 vector. pCIT767 vector was isolated from *E. coli* cells, using Qiagen plasmid isolation protocol, and used as a template for the Polymerase Chain Reaction (PCR).

For verification of *GPA1* sequence, pGEM-T Easy+*GPA1* construct was also isolated from *E. coli* JM109 cells using the same system. Restriction enzyme analysis

was carried out on the isolated construct using EcoRI restriction enzyme. Sequencing was carried out at Seqlab GmBH (Göttingen, GERMANY).

5.5 Homology Modeling

Throughout this study, fragment-based homology modeling method was used. *GPA1* sequence (GenBank entry: M32887), from *Arabidopsis thaliana*, was processed against a database of PDB files with known structures using PSIBLAST (Altschul SF *et al.*, 1997). The most similar 10 sequences were downloaded onto WORKBENCH (*http://workbench.sdsc.edu*) and an alignment was generated with CLUSTALW (Thompson JD *et al.*, 1994). After inspection of the result, the number of sequences for final alignment was reduced to two, namely the sequences 1fgk (Chimera of guanine nucleotide-binding protein G_t α-1 subunit and guanine nucleotide-binding protein G_i α-1 subunit) and 1gg2 (G protein heterotrimer mutant G_i α-1, β-1, γ-2 with GDP bound) in PDB that are rat transducin α subunits.

Final alignment was loaded onto the MODELLER program within InsightII, to create two models with low loop optimizations. The accuracy of the models was checked using ERRAT (Colovos C and Yeates TO, 1993), a program for verifying protein structures determined by crystallography. Here the error values are plotted as a function of the position of a sliding 9-residue window. The error function is based on the statistics of non-bonded atom-atom interactions in the reported structure (compared to a database of reliable high-resolution structures). An acceptable structural model should have 95% of its overall error function below the 95% limit, shown in ERRAT graphs.

A second alignment was done, using Align2D function within MODELLER, to generate a new set of models. These models were also checked using ERRAT. A final modeling job was carried out using the first alignment, but this time with high loop optimizations and sent to ERRAT to verify their accuracy.

6 RESULTS

6.1 Amplification of *GPA1* and Sequence Verification

pCIT767 plasmid, bearing the *GPA1* sequence was kindly provided by Dr. Hong Ma (Penn. State Univ., Pennsylvania, USA), in *E. coli* cells and served as a starting point for this study. The plasmid was isolated from the cells and digested using EcoRI restriction enzyme, to obtain a 1100 bp fragment containing *GPA1*. This fragment (Figure 6.1A) was used as a template in the PCR for amplification (Figure 6.1B) with internal primers designed according to GPA1 sequence. As shown in figure 6.1B, the PCR product is a fragment of about 1100 bp.

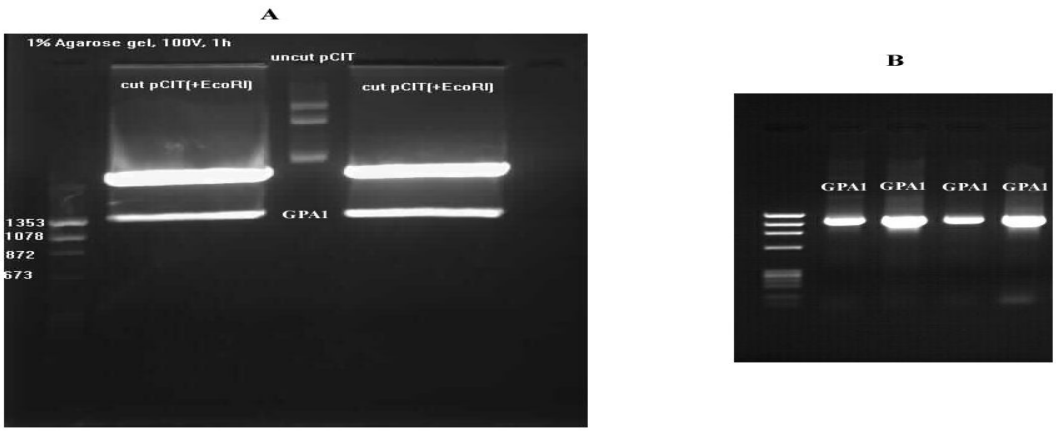


Figure 6.1 A) Results of agarose gel electrophoresis showing pCIT767 vector cut with EcoRI restriction enzyme. The first lane is the MW marker, Φ174/BSU (Promega). Second, third and fifth lanes are digested pCIT767 showing *GPA1*. The fourth lane shows uncut pCIT767. B) Agarose gel electrophoresis results showing PCR products using *GPA1*, cut from pCIT767, as a template. Lane 1 is the marker and lanes 2-5 are PCR products.

The amplified fragment was ligated into pGEM-T Easy vector and *E. coli* JM109 cells were transformed with the construct. Plasmid isolation and restriction enzyme analysis, as well as sequencing of the pGEM-T Easy+*GPA1* construct verified that *GPA1* was amplified without any mismatches (Figure 6.2). The gene is being currently cloned for over-expression.

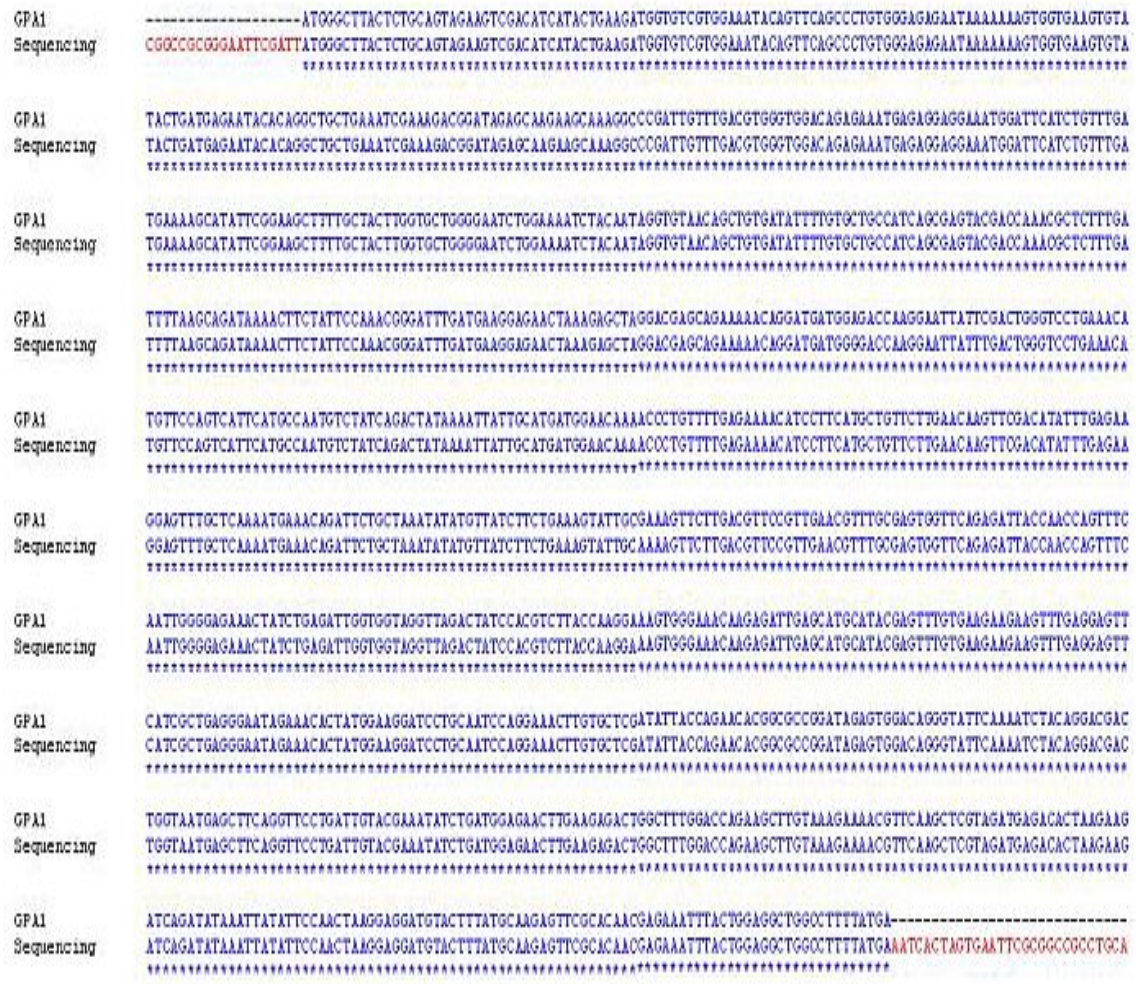


Figure 6.2 Sequencing results for pGEM-T Easy+*GPA1* construct isolated from E. coli JM109 cells. Identical residues are shown in blue, while pGEM-T Easy sequence is shown in red.

6.2 Structure Prediction of GPA1

6.2.1 Alignment of GPA1 with Mammalian Gα

The amino acid sequence for G-protein α subunit (AAA32805) from *Arabidopsis thaliana* (GPA1) was searched against proteins of known structure, using PSIBLAST (Altschul SF *et al.*, 1997; Schaffer AA *et al.*, 2001). (Appendix A) Highest score yielding proteins were found to be members of the transducin family, especially of rat (*Rattus norvegicus*) origin. Out of these sequences, the ten most similar were aligned amongst themselves using CLUSTALW (Thompson JD *et al.*, 1994). As can be seen in Figure 6.3, there is 62% identity between mammalian sequences. Alignment of GPA1 sequence with the rat sequences using CLUSTALWPROF is shown in Figure 6.4. Similar protein sequences were removed from the rat transducin alignment and a final alignment, with only two transducin sequences and the GPA1 sequence, was generated. As demonstrated on Figure 6.5, the identity among the sequences decreased to 40%.

6.2.2 Secondary Structure Prediction of GPA1

Secondary structure prediction was carried out by sending the GPA1 sequence to the META server (*http://www1.embl-heidelberg.de/predictprotein/submit_meta.html*) that uses three different prediction servers, PSIpred (Jones DT, 1999), PSSP (Raghava G, unpublished) and SAM-T99 (Karplus K *et al.*, 1998). The resulting predictions were compared using a program written in PERL (Sezerman U, unpublished) and the consensus secondary structure shown in Figure 6.6 was obtained. Alignment of the predicted structure with GPA1 and rat transducin is shown in Figure 6.7. As can be seen from these comparisons and Figure 6.8, the key residues are conserved across mammalian and plant species, as well as between plant species with the exception of AC-binding site. The overall predicted secondary structure agrees 82% with the observed structure of rat transducin α, excluding the gapped regions found in the

alignment. The GTP-hydrolysis region (res. 40-54) as well as switch I (res. 189-199), II (res. 216-237) and III (res. 249-260) are almost fully conserved across species.

6.2.3 Secondary Structure Prediction of β and γ subunits

Same procedures for secondary structure prediction were carried out for G-protein β and γ subunits. Both subunits were aligned with the most similar sequences from PSIBLAST results, using CLUSTALW followed by CLUSTALWPROF. Each of the subunits was sent to the META server for secondary structure prediction. The alignment results are shown in figures 6.9 and 6.10 together with those of secondary structure predictions. Results on figure 6.9 indicate that *A. thaliana* β subunit sequence agrees 51% with that of human transducin β subunit whereas those on figure 6.10 show γ subunit with the bovine counterpart. These results are consistent with those obtained for the α subunit that the best agreement is within the transducin family. Among the known plant G-protein β subunit sequences, key residues for binding α and γ subunits were found to be mostly conserved (Figure 6.11), while the γ subunit lacked the conservation of key residues.

6.2.4 Modeling the GPA1 structure

3D structural models were generated, using the CLUSTALW alignment of GPA1 with two rat transducin α subunits (PDB entries: 1GG2 and 1FQK) and the Align2D alignment module within the MODELLER program.

The first models were created using low loop optimizations within the program. These models had up to 62% of their structure below the given error limits when sent to ERRAT (Colovos C and Yeates TO, 1993), and yielded erroneous folding as can be seen from the Figures 6.12 A and B.

Align2D function within the MODELLER program was used to generate a new alignment with the same sequences. This time the amount of correctly folded regions decreased down to 26%. (Figure 6.13)

Two new models with high loop optimizations using CLUSTALW alignment were created. These yielded 76% of the structure within the limits. The model shown on figure 6.14-B was better folded at the amino-terminus (res. 1-10) and at different positions corresponding to several functional sites, i.e. GTP-hydrolysis region (res. 40-45), switch I (res. 193-203), switch III (res. 256-259). As can be seen by comparing figure 6.14-B with figure 2.3 the two-domain structure of the α subunit can be identified. Appearance of the GTP-binding pocket and possible distribution of helices and extended structures (based on figures 6.6 and 6.7) indicate that the loop regions within the created models do not interfere with any functional sites, but alter the overall structure of GPA1.

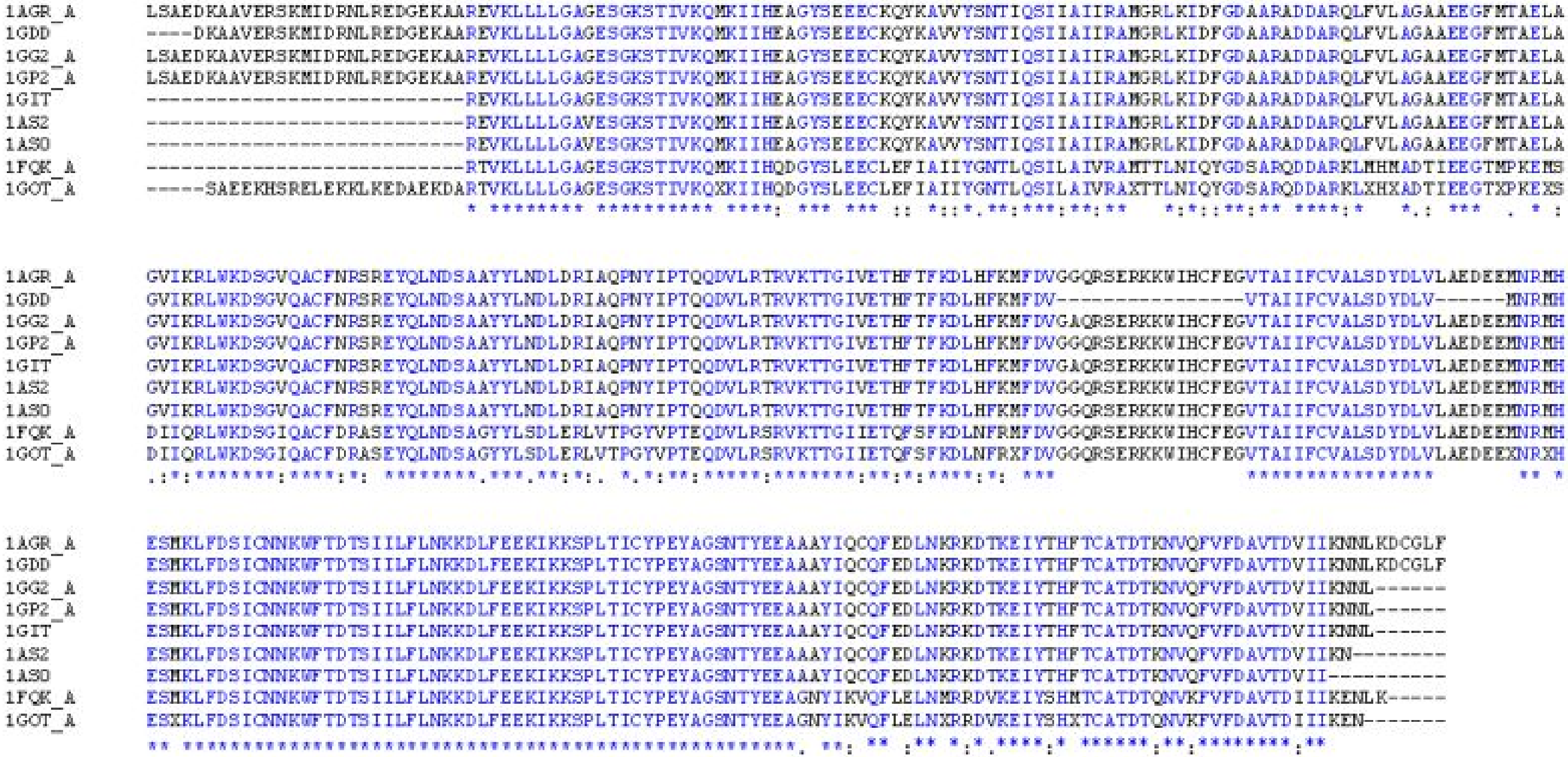


Figure 6.3 Alignment of rat transducin α protein sequences using PSIBLAST. These were found to be those that are most similar to GPA1. PDB codes are indicated on the left. Identical amino acids are shown in blue, with `*`. `: ` indicates strong similarity while `.` indicates low similarity between amino acids.

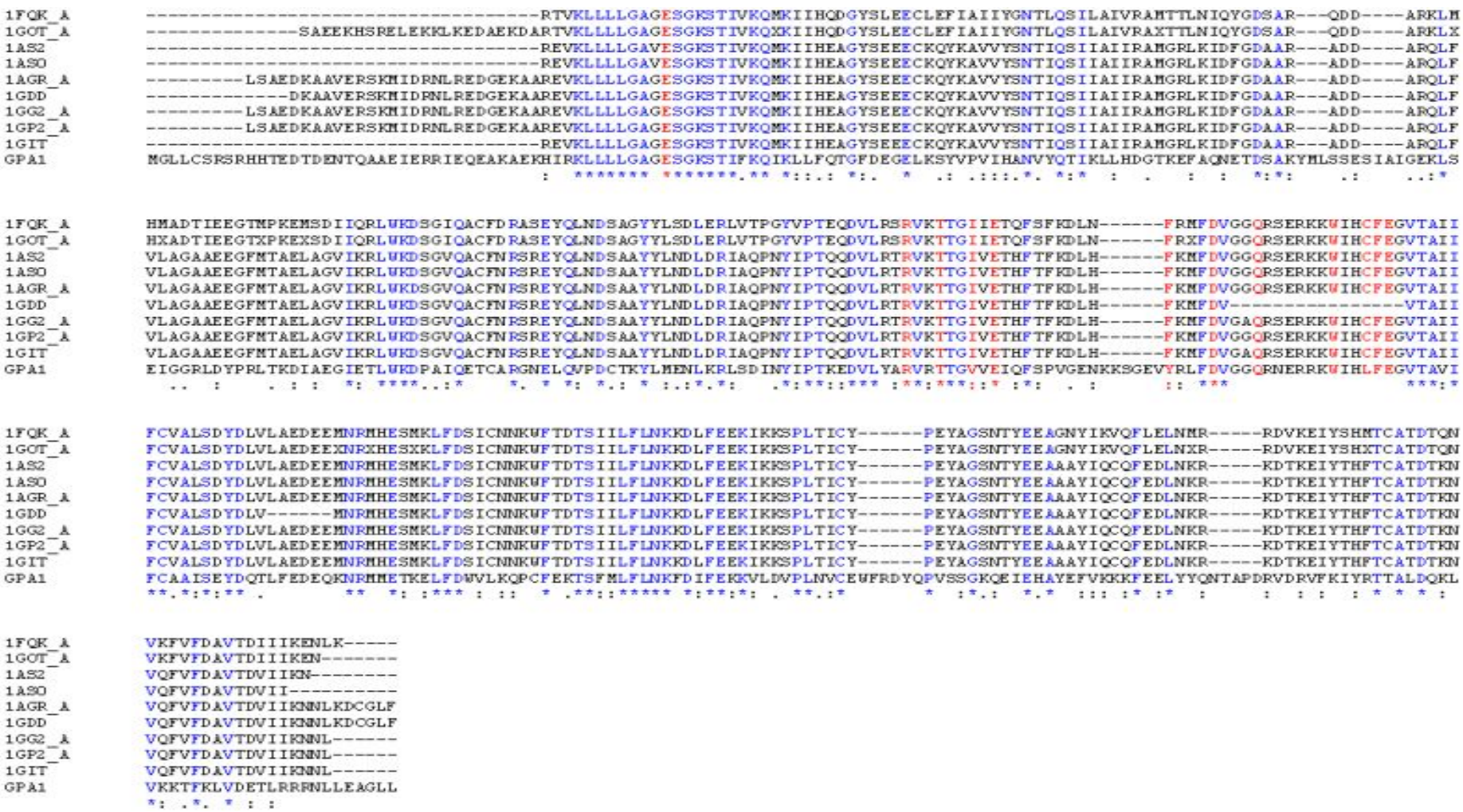


Figure 6.4 CLUSTALW PROF alignment of GPA1 with the most similar rat transducin α sequences.

Figure 6.5 CLUSTALWPROF alignment generated using the two most similar rat transducins and GPA1 sequences.

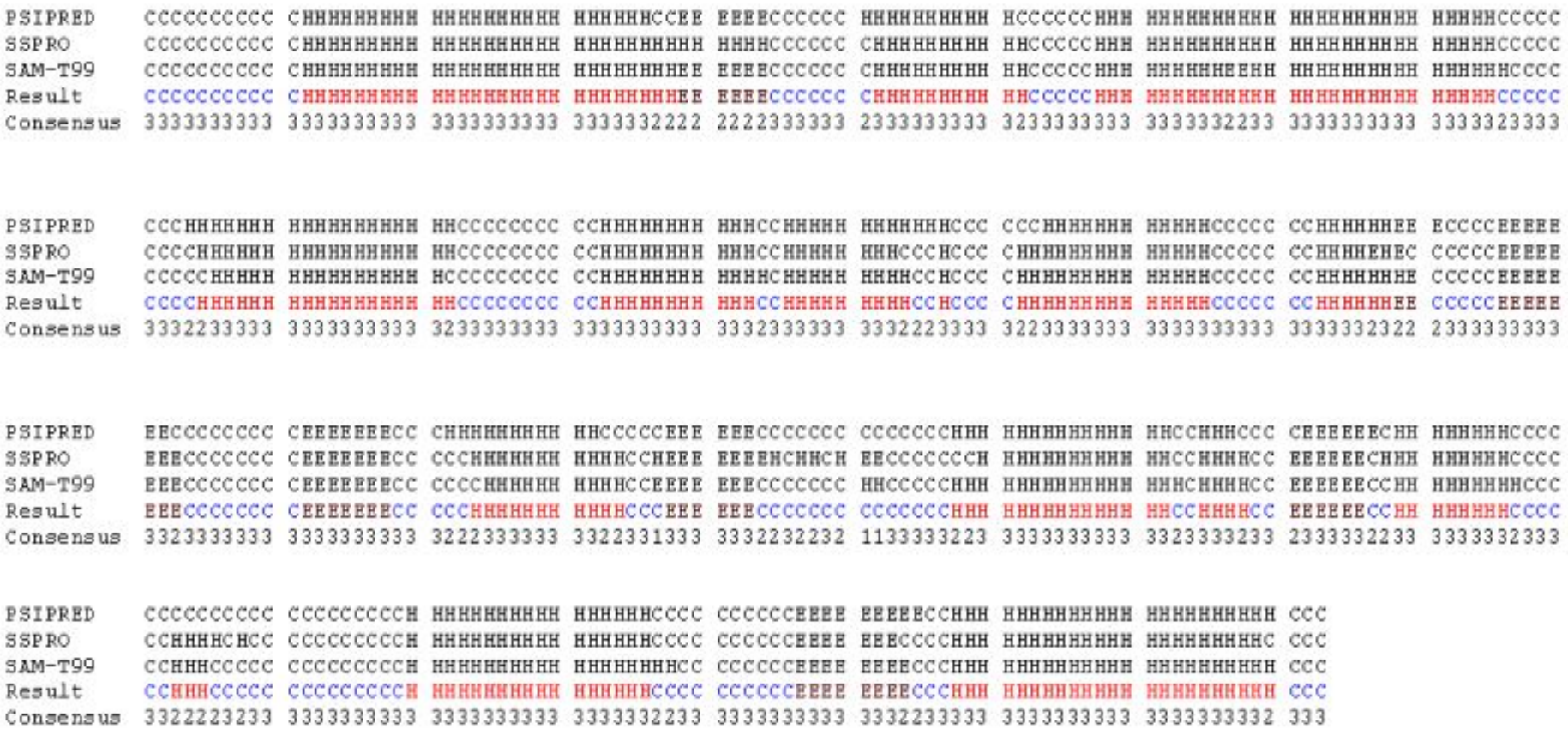


Figure 6.6 Secondary structure prediction results for GPA1. PSIpred, SSPRO and SAM-T99 are the three structure predicting server names and the colored sequence is the resulting prediction (C=Coil, H=Helix, E=Extended).

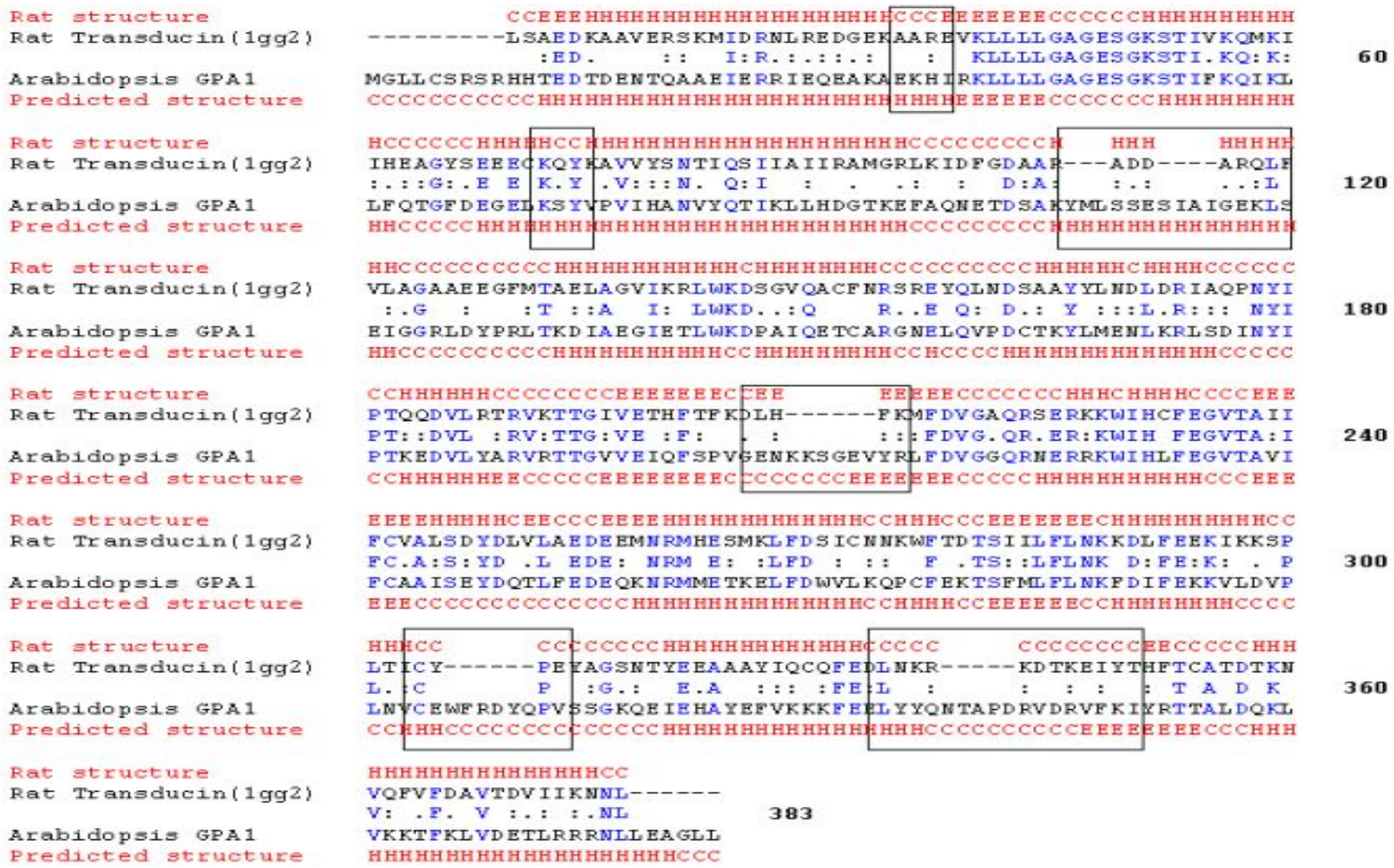


Figure 6.7 Sequence and secondary structure alignment of rat transducin and GPA1 using CLUSTALW, where secondary structure features are shown in red. Identical amino acids are shown in blue and similar amino acids in green. Boxes indicate erroneous predictions.

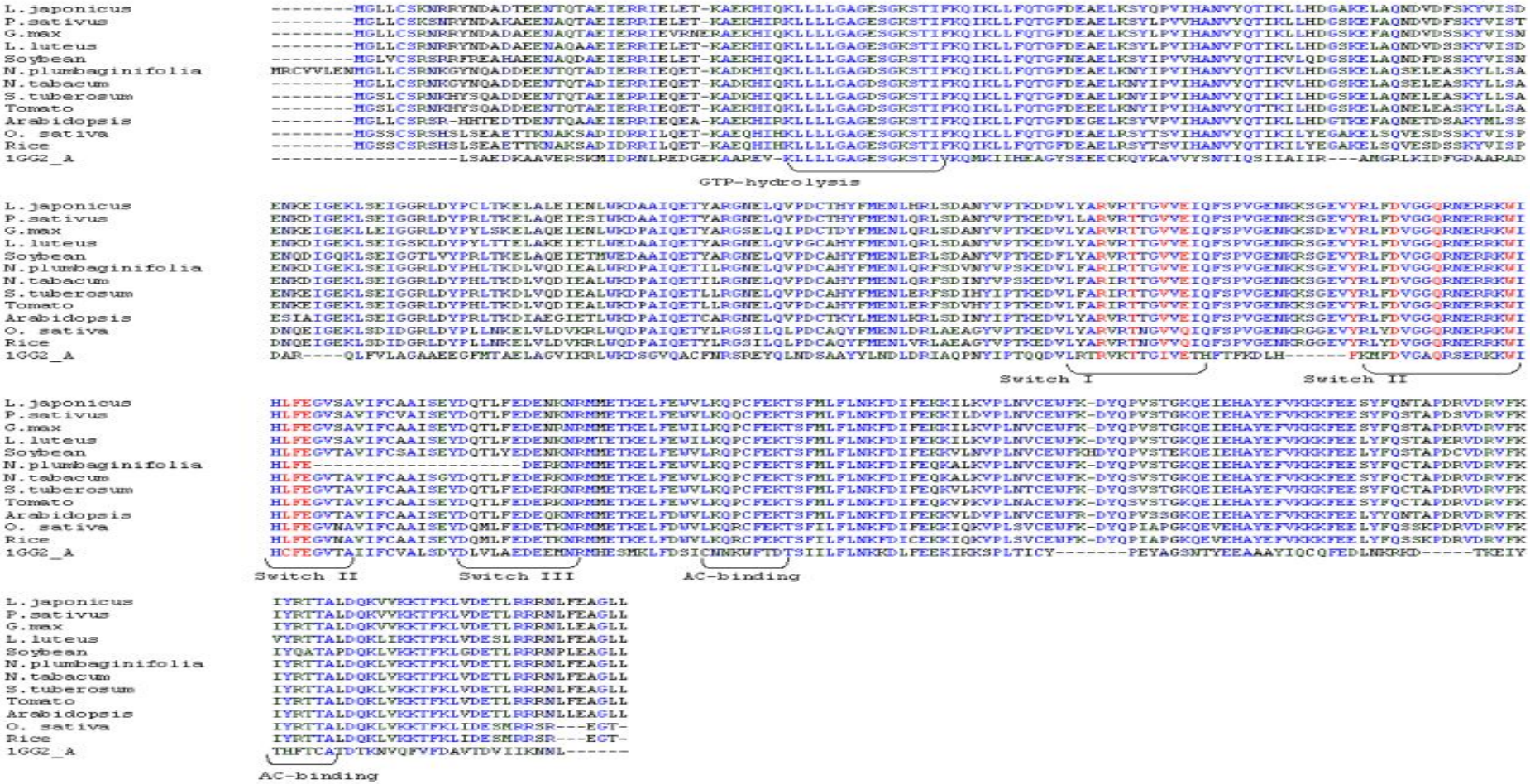


Figure 6.8 CLUSTALW Alignment of all known plant G-protein α subunits and rat transducin α protein 1GG2. Residues in blue are identical, highly similar ones are in green and key residues are shown in red.

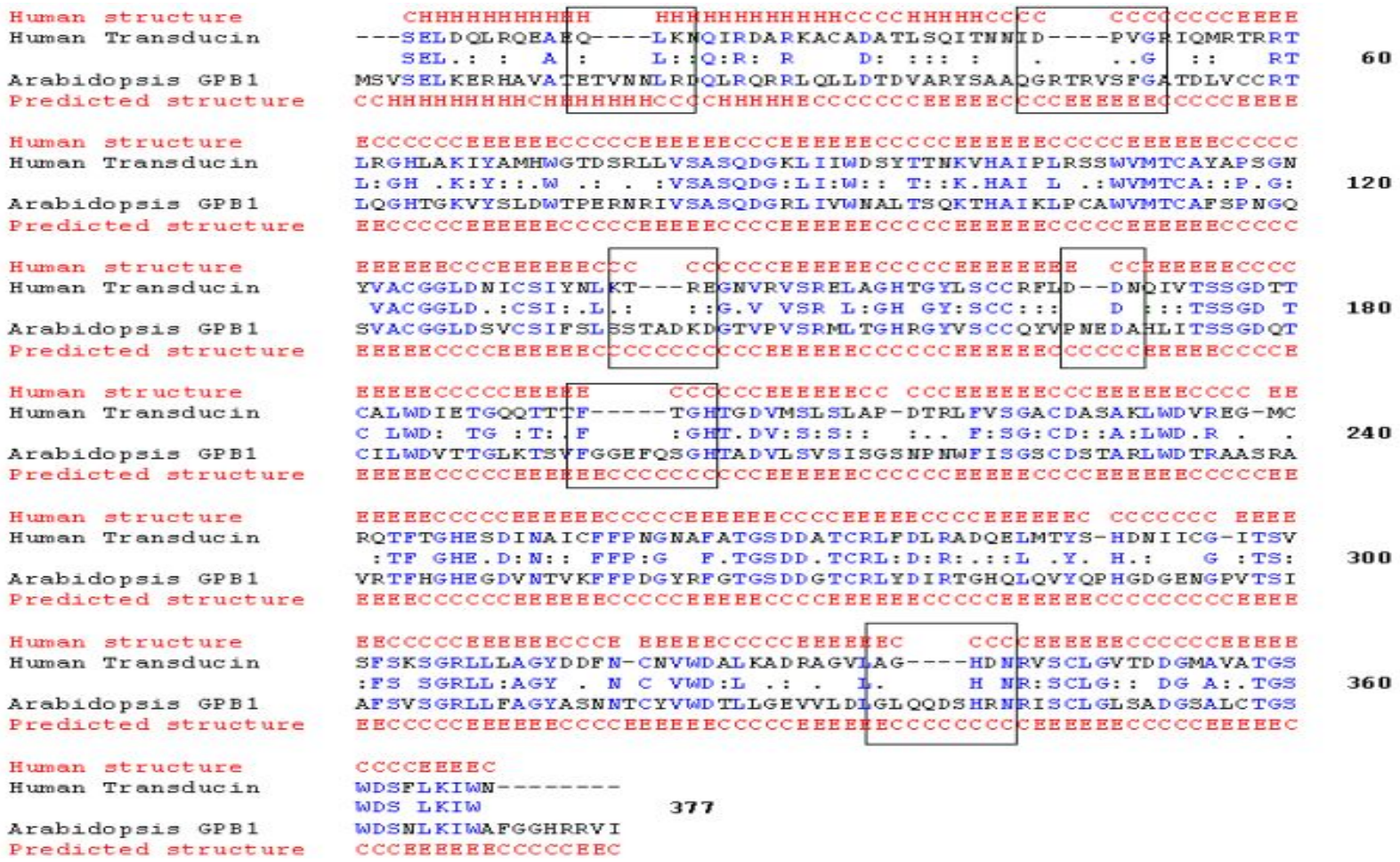


Figure 6.9 G-protein β subunit (GBB1) from A. thaliana aligned with human transducin β subunit (1GG2_B) using CLUSTALW. Secondary structure prediction results are shown in red. (H=Helix, C=Coil, E=Extended) Boxes indicate erroneous predictions.

A

```
Bovine structure                  CHHHHH    HHHHHHH HHHCCCCCCHHHHHHH     HHHHHHH
Bovine transducin       ----------SIAQAR---KLVEQLK-MEANIDRIKVSKAAAD----LMAYCEA----
                                  S::::.    ::: :L  :E ::  :: .   .:     : : CE        60
Arabidopsis gamma AGG1  MREETVVYEQEESVSHGGGKHRILAELARVEQEVAFLEKELKEVENTDIVSTVCEELLSV
Predicted structure     CCCCCEEEECCHHCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCHHHHHHHHHH

Bovine structure         HHHHCCCCC    CCCCCCCCC
Bovine transducin       -HAKEDPLLT----PVPASENPF---------------
                         .   DPLL.    P:  . : :                  98
Arabidopsis gamma AGG1  IEKGPDPLLPLTNGPLNLGWDRWFEGPNGGEGCRCLIL
Predicted structure     HHCCCCCCCCCCCCCCCCCCHHCCCCCCCCCCCEEEC
```

B

```
Bovine structure                    CHHHHH   HHHHHHH         HHHCCCCCC HHHHHHHHHHHH
Bovine transducin       ------------SIAQAR---KLVEQLK--------MEANIDRIK-VSKAAADLMAYC
                                    :.::R   ::   :LK         :E :::::: :.:A:A.   :    60
Arabidopsis gamma AGG2  MEAGSSNSSGQLSGRVVDTRGKHRIQAELKRLEQEARFLEEELEQLEKMDNASASCKEFL
Predicted structure     CCCCCCCCCCCCCCCEEECCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHCHHHHHHHHH

Bovine structure        HHH HHHCCCC    CCCCCCCCCC
Bovine transducin       EAH-AKEDPLL----TPVPASENPF---------------
                        ::   :K DPLL     PV A: : :                 100
Arabidopsis gamma AGG2  DSVDSKPDPLLPETTGPVNATWDQWFEGPKEAKRCGCSIL
Predicted structure     HHHHCCCCCCCCCCCCCCCCCCHHHCCCCCCCCCCCCCCC
```

Figure 6.10 G-protein γ subunits (AGG1 and AGG2) from A. thaliana aligned with bovine transducin γ (1GG2_G) using CLUSTALW. Secondary structure prediction results are shown in red.

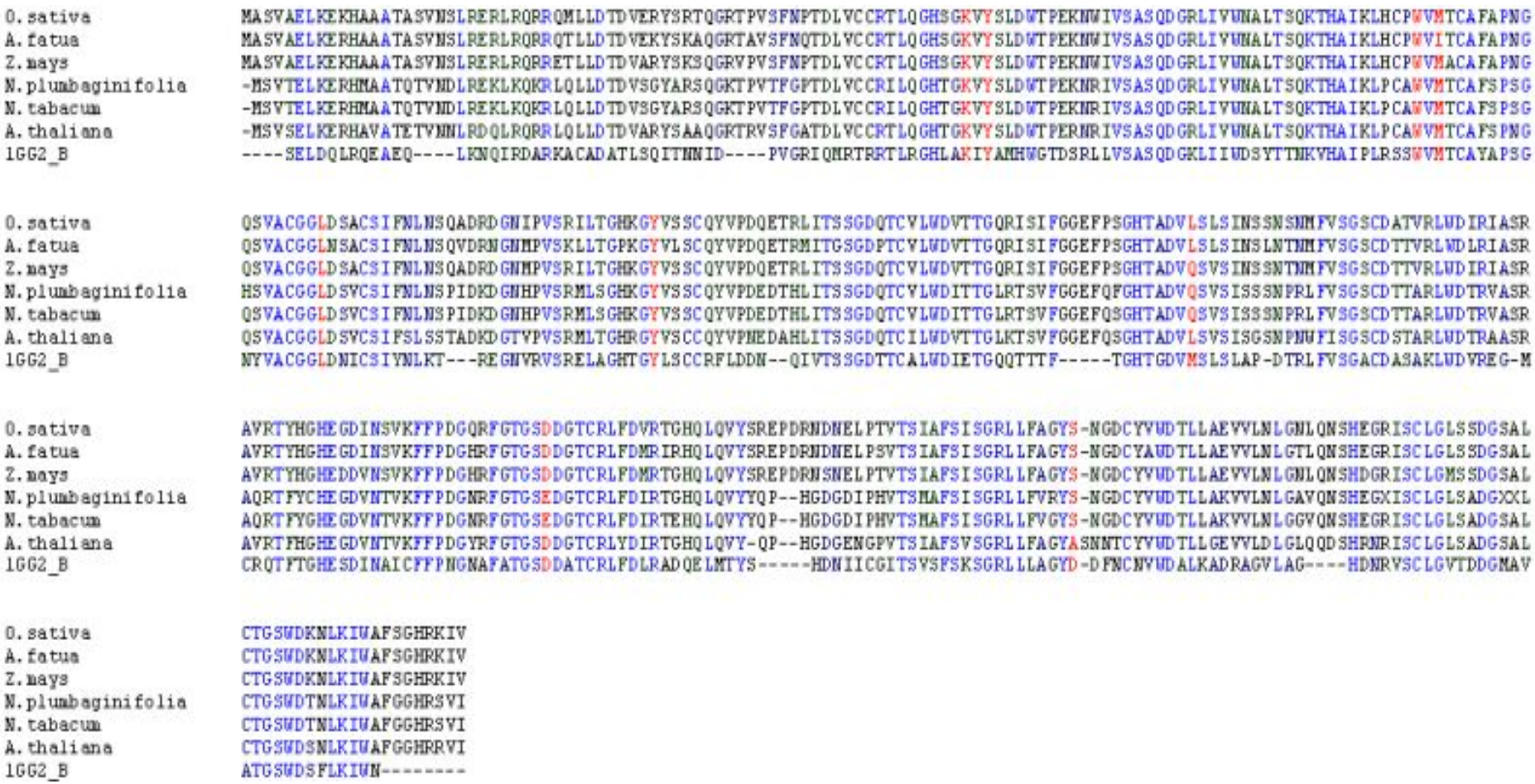
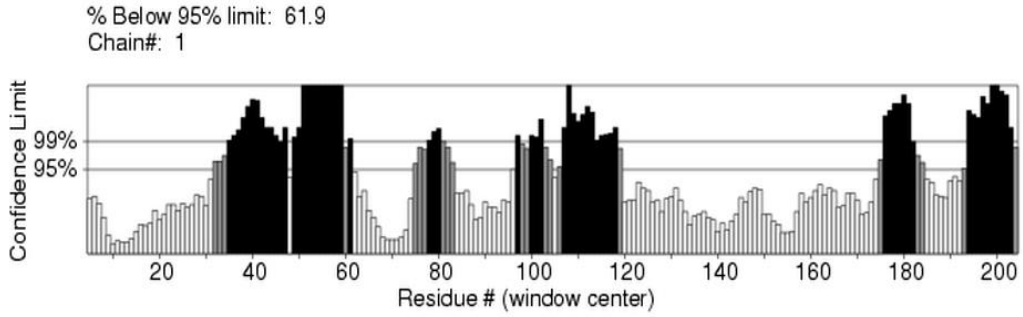
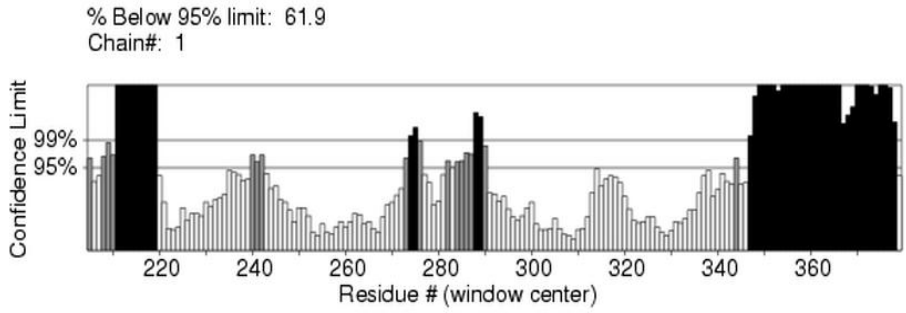


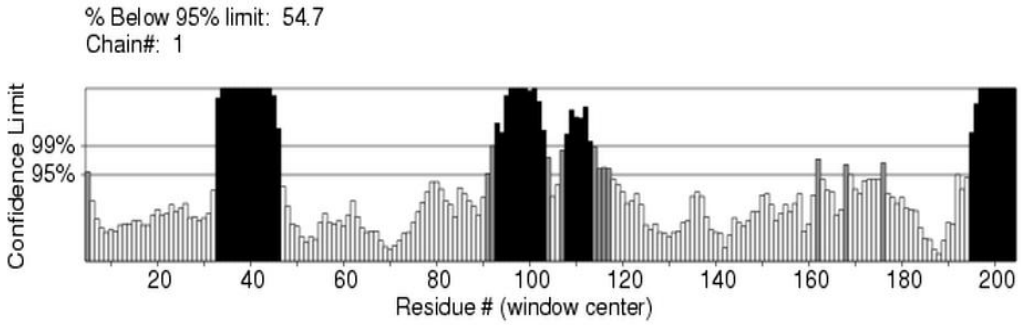
Figure 6.11 CLUSTALWPROF Alignment of all known plant G-protein β subunits with the most similar mammalian sequence, bovine transducin (1GG2_B). Identical residues are shown in blue, similar sequences are shown in green. Key residues in binding α subunit are shown in red.

A


% Below 95% limit: 61.9
Chain#: 1
Confidence Limit
99%
95%
20
40
60
80
100
120
140
160
180
200
Residue # (window center)


% Below 95% limit: 61.9
Chain#: 1
Confidence Limit
99%
95%
220
240
260
280
300
320
340
360
Residue # (window center)

B



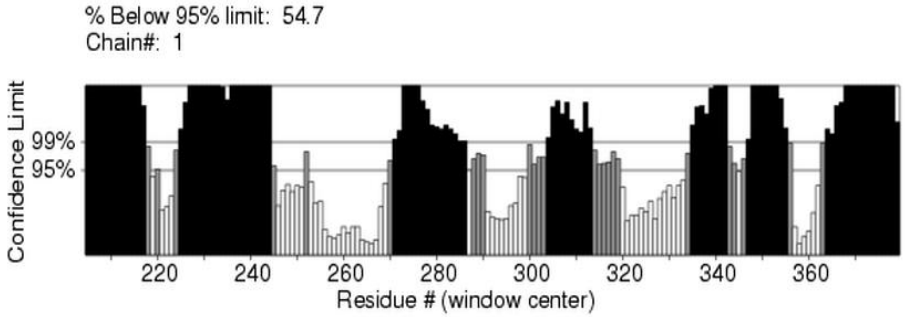
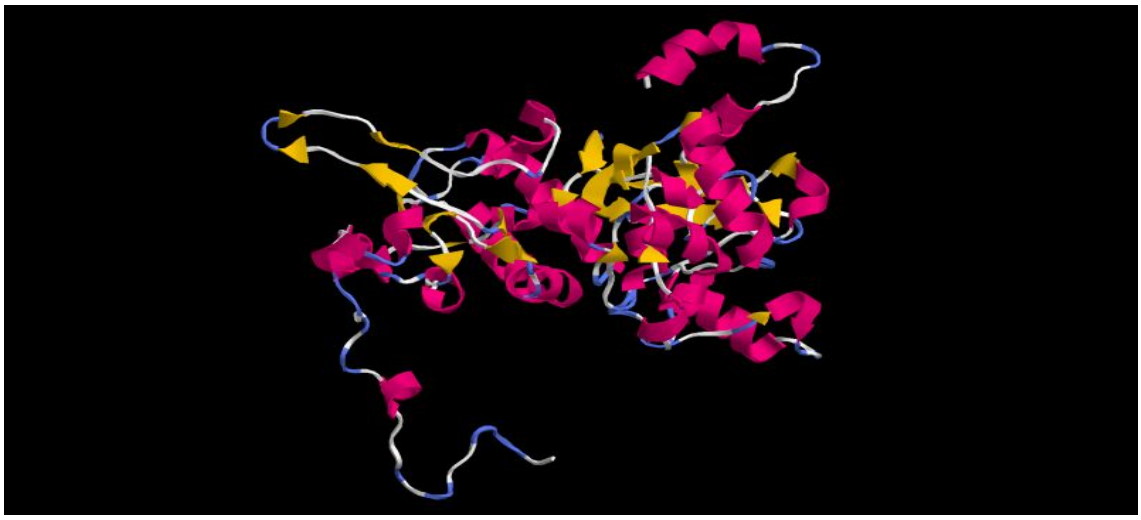
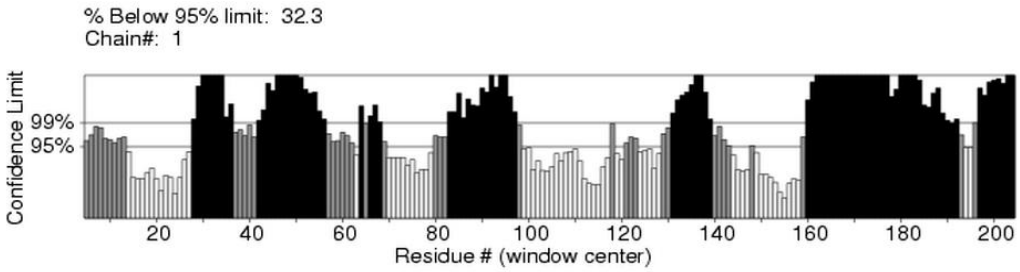
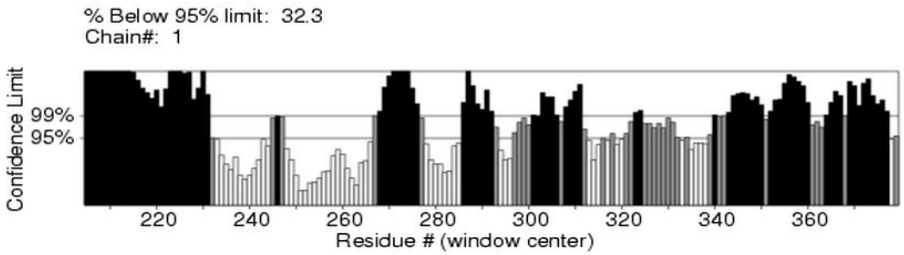


Figure 6.12 G-protein α subunit models created using MODELLER, with low optimization parameters. A) Model 1 with the ERRAT result graphs. B) Model 2 with the ERRAT result graphs.

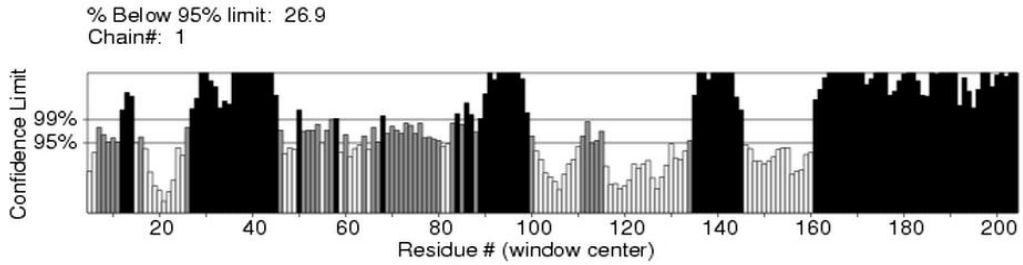
A




% Below 95% limit: 32.3
Chain#: 1
Confidence Limit
99%
95%
20
40
60
80
100
120
140
160
180
200
Residue # (window center)


% Below 95% limit: 32.3
Chain#: 1
Confidence Limit
99%
95%
220
240
260
280
300
320
340
360
Residue # (window center)

B



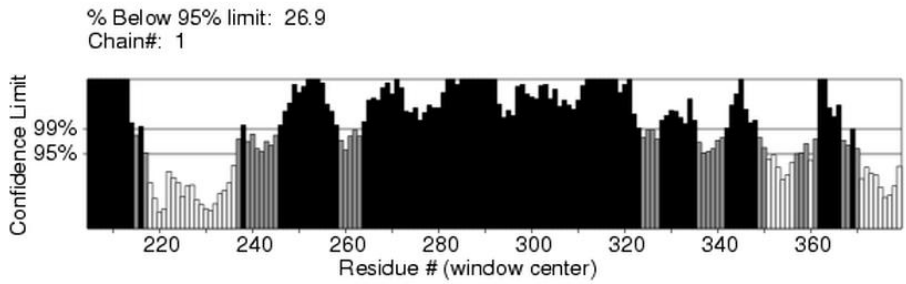
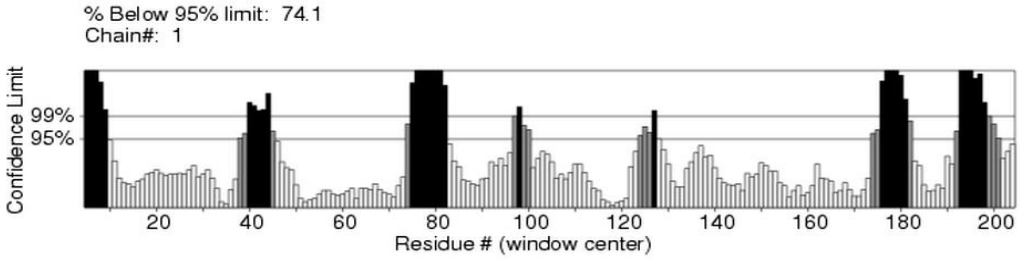
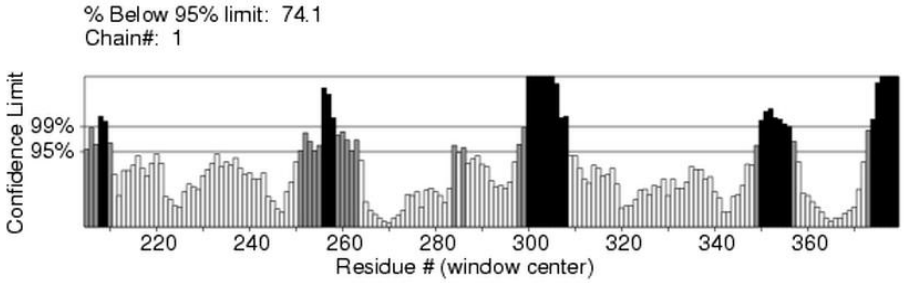


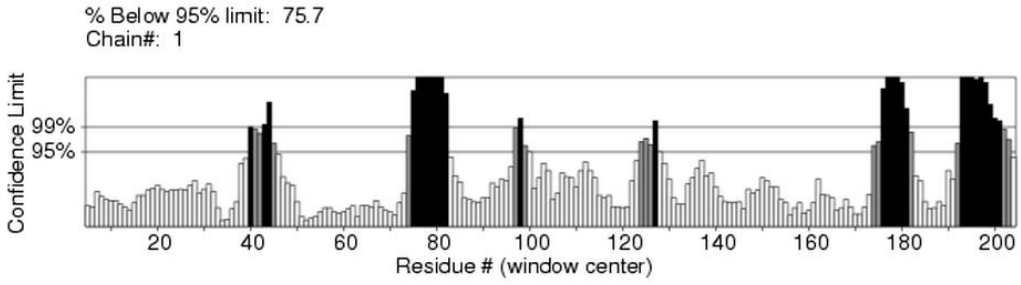
Figure 6.13 Models created using Align2D function within MODELLER. A) Model 1 created with its ERRAT result. B) Model 2 with its ERRAT result.

A


% Below 95% limit: 74.1
Chain#: 1
Confidence Limit
99%
95%
20
40
60
80
100
120
140
160
180
200
Residue # (window center)


% Below 95% limit: 74.1
Chain#: 1
Confidence Limit
99%
95%
220
240
260
280
300
320
340
360
Residue # (window center)

B



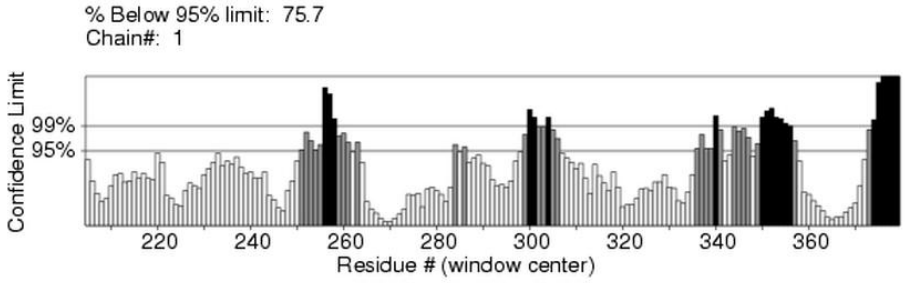


Figure 6.14 Models created using CLUSTALW with high loop optimization. A) First model with its ERRAT result. B) Second model with its ERRAT result.

7 DISCUSSION

Biological macromolecules such as proteins and nucleic acids perform crucial tasks that sustain life. Specific tasks, such as catalysis of an enzymatic reaction or a ligand-binding event, are intimately associated with the structure and the structure of the molecule may undergo conformational changes while performing its function. Defining the molecular structure and its dynamic behavior is, therefore, essential for understanding how a biological macromolecule functions.

Current structural studies use two different approaches: Experimental and computational methods. The experimental approach involves isolation of the molecule directly or cloning, expression and purification of the protein of interest followed by NMR (Nuclear magnetic Resonance) or X-Ray crystallography experiments to determine the 3D structure. The experimental approach is not only time consuming but both X-Ray and NMR methods have limitations and need to be refined for yielding true structures.

In X-Ray crystallography, the interaction of X-Rays with electrons is used to obtain an electron-density map of the molecule, which can then be interpreted in terms of an atomic model. Crystallization of proteins can be difficult to achieve and usually requires many different experiments varying a number of parameters, such as pH, temperature, protein concentration, and the nature of solvent and precipitant. Protein crystals contain large channels and holes filled with solvents, which can be used for diffusion of heavy metals into the crystals. The addition of heavy metals is necessary for the phase determination of the diffracted beams. X-Ray structures are determined at different levels of resolution. At low resolution, only the shape of the molecule is obtained, whereas at high resolution most atomic positions can be determined to a high degree of accuracy. At medium resolution, the fold of the polypeptide chain is generally

correctly revealed together with the approximate positions of the side chains, including those at the active site (Branden C and Tooze J, 1999).

In NMR, the magnetic-spin properties of atomic nuclei within a molecule are used to obtain a list of distance constraints between those atoms in the molecule, from which a three-dimensional structure of the protein molecule can be obtained. However, not requiring any protein crystals, this method is restricted in its use to small protein molecules (Cantor CR and Schimmel PR, 1980).

Computational methods, on the other hand, require only the sequence of the protein in question as the input, but different techniques for comparative modeling, including homology modeling, threading as modeling protein-ligand complexes and protein-protein complexes need to be developed. Computational approaches depend on the algorithms created and the hardware used; they try to correlate biological function with structure while creating a model of the protein of interest. As there are no considerable limitations for the use of these techniques, they are very fast yet less reliable than experimental procedures.

In this study, preliminary results of verification of *GPA1* sequence and homology-modeling results for predicting the 3D structure of *Arabidopsis thaliana* G-protein α subunit are presented. As there are no experimental or calculated models reported for GPA1, these results provide insight for functional analyses and a basis for further structural studies.

7.1 Subcloning and *GPA1* Sequence Verification

PCR amplification of *GPA1* and further subcloning studies, using pGEM-T Easy, showed that the amplified *GPA1* sequence agreed with that published by Ma, H (Ma H *et al.*, 1990). This result provided a basis for further cloning work for over-expression of *GPA1*. Over-expressed recombinant GPA1 will be purified and used in X-Ray

crystallography and solution scattering studies for determination of the structure experimentally.

7.2 Structure Prediction and Modeling

Despite extensive work on the structure of mammalian heterotrimeric G-proteins and interpretation of their functional characteristics in terms of structural features (Sprang SR, 1997) similar studies on plant G-proteins are lacking (Fujisawa Y *et al.*, 2001). It is still uncertain if G-proteins function as heterotrimers in plants, which signalling pathways they are involved in and if the mechanisms of activation (which involves interactions with GPCRs) and action (which involve interactions with effectors and regulators) are similar to the mammalian systems. Lack of direct experimental structural data renders sequence analysis, structure prediction and modelling studies such as those reported in this thesis useful for providing clues to answers to these questions. The results presented in this work can be interpreted in the framework of functional sites and domains and their relation to mechanism of activation and action.

Alignment operations indicated that functional sites and regions such as β, γ and GTP binding, are conserved between GPA1 and rat transducin (Figure 6.8). Key residues in βγ binding i.e. Arg-178, Ile-184, Glu-186, Phe-199, Trp-211, Cys-214, Phe-215 and Glu-216 are mostly conserved with the exception of Ile-184, Phe-199, and Cys-214 that are mutated to Val, Tyr, and Leu in GPA1. As mentioned earlier GTP-hydrolysis (KLLLGAGESGKST) region is also fully conserved as well as the amino acids Gln-204 and the Thr-181 involved in GTP- and Mg^{+2} binding respectively.

Switch regions (I, II and III) important in binding the βγ subunit and stabilizing the GTP-bound state are found to be highly homologous. Here differences appear to be substitutions from similar groups of amino acids i.e. Thr→Ala, Lys→Arg.

These results indicate that the mechanism of GTP-binding and hydrolysis by G-proteins in plants appear similar to that observed in mammalian species. These studies

are further supported by biochemical work on GTP binding and GTPase activity (Perroud PF *et al.*, 1997).

A drastic difference in protein sequence between rat transducin and GPA1 is observed in the region corresponding to adenylyl cyclase binding site in transducin. As can be seen on figure 6.8, this region is fully conserved among the plant species. This result implies that plant G-proteins are not involved in the AC pathway in plants at least not at this site. Moreover as AC is not yet found in plants it may point to different mechanisms for cAMP mechanisms in plants. Since AC is an effector of G-proteins in mammalian systems the result presented here may also point to different effector interactions.

Alignments of the known 14 plant G-protein α subunit sequences demonstrate the high level of homology (70-85%) among plant species. As expected key residues in βγ-binding, GTP-hydrolysis and the switch regions are highly conserved and indicate similar mode of interaction with the β subunit and similar functional features. (Figure 6.8)

As mentioned earlier, in mammalian systems heterotrimeric G-protein specificity is established mostly by the α subunit which appears to show sequence variations according to the tissue in different plants. Based on sequence analysis only two types of G α subunit is reported in the Arabidopsis genome. One is GPA1 and the other is an extra large G α to which no function has so far been attributed (Lee YR and Assmann SM, 1999). The small number and the high homology among the plant G α subunits may indicate that there are a small number of signalling pathways where G-proteins are involved.

Secondary structure predictions yielded a high degree of identity (95%) between the structures of rat transducin (1gg2) and GPA1. The differences occurred due to gaps within the alignment and resulted in longer helices or coils in GPA1 structure. In order to overcome this problem, the alignments can be altered by changing the parameters (Appendix A) or the positions of the amino acids within the alignment, manually.

All plant G protein α subunits are members of the inhibitory type of G protein family, G_i. They show the highest sequence similarity to transducin proteins within this family, but have the characteristic functional sites of G_z proteins. Both G_z proteins and plant G protein α subunits lack the carboxyl-terminal cysteine required for ribosylation by pertussis toxin and contain a myristolization motif specific to G_z proteins. As homology modeling methods require the usage of sequences with the highest identities, transducin α subunits were used in our studies. In order to preserve the functional sites, G_z proteins could be used instead of transducins to generate models.

Results of alignment and secondary structure prediction for β subunits support interaction of α and β subunits in plants. This provides evidence that G-proteins may function as heterotrimers in plants. Similarities observed in the plant β subunit sequences with their mammalian counterparts (figures 6.9 and 6.10) complement the results for the α subunit, and point towards a similar mechanism of activation and action in plants.

Plant γ subunits, on the other hand, lack the key residues required for β binding (amino- and carboxyl-termini), and have very little similarity with the mammalian γ subunits. (Figure 6.11) There are currently 2 known γ subunits in plants, identified in *A. thaliana*, having similarities up to 40% with each other. Using the positively charged amino acids at the amino-terminus of the γ subunit, it is possible to find candidate sites on the α subunit for binding.

Studies conducted on βγ binding in plants, revealed two different heterotrimer formations, depending on the γ subunit involved (Mason MG *et al.*, 2001). It is possible that the specificity of the G-proteins in plants is provided by the γ subunit.

Interpretation of sequence alignment results in terms of mechanism of activation and action of plant G-proteins is further supported by the results of secondary structure prediction work. The predicted structure was found to be very similar to the known structure of rat transducin, except in some helix and loop regions. Long helices in GPA1 structure are found after the GTP-hydrolysis site and at the C-terminus, both not affecting the function of the subunit directly. Although not affecting the function

directly, these longer helices change the overall conformation of the protein, resulting in a different structure.

For modelling GPA1, homology modelling methods were used. Although being the best method available, homology modelling methods depend on the parameters of the alignments that are used. In turn, altering the parameters to improve the alignments requires expertise and is the weakest part of this work. As the alignments used in this study relied on sequence conservations, more than functional conservations, further work on the alignments is required.

The structural model developed for the Arabidopsis G α subunit results in a two domain structure similar to that of rat transducin as can be seen on figure 6.14 B. The GTP binding and hydrolysis pocket is identified. This model needs to be improved by further work on the loop regions. This is a task where currently there are no effective methods (Fiser A *et al.*, 2000). Although the developed model was sent to ERRAT for verification purposes, the results of ERRAT can not be used solely as ERRAT checks the structure regardless of the position of the amino acids. Thus, other programs need to be used to verify the accuracy of the model.

8 CONCLUSION

G-proteins are important constituents of cellular signaling, having roles in various processes ranging from development and diseases in mammalians to pathogen resistance and hormone signaling in plants. Defining their structure will provide insights into distinct processes among different species as defined above. Mammalian G-proteins have been studied extensively, together with the mechanisms of the signal transduction cascades they are involved in. There is little knowledge on plant G-proteins and their signal transduction cascades and no structural work, which can be interpreted in the context of functional mechanisms.

With this aim, G-protein α subunit (GPA1) from *A. thaliana* has been cloned for further studies involving expression, purification and structural analysis using small angle X-Ray solution scattering and X-ray crystallography. Sequence alignment, secondary structure prediction and 3D structure prediction tools have been used to predict the structure of GPA1and to gain insights on mechanisms of signal transduction in plants.

G-protein α subunit, being the key subunit in the heterotrimer, is involved in activation of the heterotrimer and signaling of downstream effectors. Binding of GTP triggers activation of the subunit, releasing it from the heterotrimer. Upon GTP-hydrolysis, the α subunit binds to the βγ subunits to form the heterotrimer and becomes inactive. To perform these processes, the α subunit has several functional sites, including GTP-hydrolysis and switch regions.

When compared with the mammalian G-protein α subunits, functional sites were found to be highly conserved in plant G-protein α subunits, with the exception being the adenylyl cyclase binding site. As no adenylyl cyclases have so far been found in plant species, it can be concluded that there are either no adenylyl cyclases in plants or they are significantly different from their mammalian counterparts.

There are currently 14 plant G-protein α subunits known, with over 70% identity in their amino acid sequences. All of the known plant G-protein α subunits have conserved GTP-hydrolysis and switch regions, and show the highest similarity with rat transducin α (Gt) proteins. The degree of homology in the GTP hydrolysis region, Mg^{+2} and GTP binding sites show that the mechanism of GTPase activity of plant Gα subunits is similar to that of rat transducin.

Despite the similarity, however, GPA1 shares interesting common features with Gz proteins belonging to the inhibitory type G-proteins (Gi). According to the data, it can be concluded that plant G-protein subunits form a new class of G-proteins, showing high similarity to transducin proteins while having common functional features with Gz proteins.

Further studies on the 3D structure of G-protein α subunits require the expression and purification of GPA1 together with the β and γ subunits, to form the heterotrimer. Reliability of the generated models can be improved with further optimizations on loop regions of GPA1. The ultimate goal is to compare the models with the experimentally determined structure.

9 REFERENCES

1. Ahmed AH, Heppel LA, "Evidence for a role of G protein βγ subunits in the enhancement of cAMP accumulation and DNA synthesis by adenosine in human cells" *J. Cell. Physiol.* 170:3 (1997) 263-271.

2. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs" *Nucl. Acids Res.* 25:17 (1997) 3389-3402.

3. Ando S, Takumi S, Ueda Y, Ueda T, Mori N, Nakamura C, "Nicotiana tabacum cDNAs encoding alpha and beta subunits of a heterotrimeric GTP-binding protein isolated from hairy root tissues" *Genes Genet. Syst.* 75 (2000) 211-221.

4. Arshavsky VY, Lamb TD, Pugh EN Jr, "G proteins and phototransduction" *Annu. Rev. Physiol.* 64 (2002) 153-187.

5. Bischoff F, Molendijk A, Rajendrakumar CS, Palme K, "GTP-binding proteins in plants" *Cell Mol. Life Sci.* 55 (1999) 233-256.

6. Blundell TL, Sibanda BL, Sternberg MJ, Thornton JM, "Knowledge-based prediction of protein structures and the design of novel molecules" *Nature* 326 (1987) 347-352.

7. Blundell TL, Carney D, Gardner S, Hayes F, Howlin B, Hubbard T, Overington J, Singh DA, Sibinda BL, Sutcliff M, "Knowledge based protein modeling and design" *Eur. J. Biochem.* 172 (1988) 513-520.

8. Bourne HR, Sanders DA, McCormick F, "The GTPase superfamily: a conserved switch for diverse cell functions" *Nature* 348:6297 (1990) 125-132.

9. Bouvier M, "Oligomerization of G-protein-coupled transmitter receptors" *Nat. Rev. Neuroscience* 2:4 (2001) 274-286.

10. Brady AE, Limbird LE, "G protein-coupled receptor interacting proteins: Emerging roles in localization and signal transduction" *Cellular Signalling* 14:4 (2002) 297-309.

11. Branden C and Tooze J, "Introduction to Protein Structure" *Garland Publishing Inc. New York* 2nd ed. (1999) 373-392.

12. Cantor CR and Schimmel PR, "Biophysical Chemistry, Part II: Techniques for the study of Biological Structure and Function" *W.H.Freeman and Company, San Francisco* 1st ed. (1980) 343-347.

13. Coleman DE, Berghuis A, Lee E, Linder ME, Gilman AG, Sprang SR, "Structures of active conformations of $G_{i\alpha 1}$ and the mechanism of GTP hydrolysis" *Science* 265 (1994) 1405-1412.

14. Colovos C, Yeates TO, "Verification of protein structures: patterns of nonbonded atomic interactions" *Protein Sci.* 2:9 (1993) 1511-1519.

15. Conklin BR, Bourne HR, "Structural elements of G_{α} subunits that interact with $G_{\beta\gamma}$, receptors, and effectors" *Cell* 73 (1993) 631-641.

16. Cook B, Bar-Yaacov M, Cohen Ben-Ami H, Goldstein RE, Paroush Z, Selinger Z, Minke B, "Phospholipase C, and termination of G-protein-mediated signalling *in vivo*" *Nat. Cell. Biol.* 2:5 (2000) 296-301.

17. Cooper DMF, Mons N, Karpen JW, “Adenylyl cyclases and the interaction between calcium and cAMP signalling” *Nature* 374 (1995) 421-424.

18. Cowburn RF, O’Neill C, Bonkale WL, Ohm TG, Fastbom J, “Receptor-G-protein signalling in Alzheimer’s disease” *Biochem. Soc. Symp.* 67 (2001) 163-175.

19. Dascal N, “Ion-channel regulation by G proteins” *Trends Endocrinol. Metab.* 12:9 (2001) 391-398.

20. Davies PJ, “Plant hormones: Physiology, biochemistry and molecular biology” *Kluwer Academic Publications* ed. (1995) Dordrecht, The Netherlands.

21. Divecha N, Irvine RF, “Phospholipid signalling” *Cell* 80 (1995) 269-278.

22. Dohlman HG, Thorner J, “RGS proteins and signaling by heterotrimeric G proteins” *J. Biol. Chem.* 272:7 (1997) 3871-3874.

23. Downes GB, Gautam N, “The G protein subunit gene families” *Genomics* 62 (1999) 544-552.

24. Farrens D, Altenbach C, Yang K, Hubbell W, Koharana H, “Requirement of rigid-body motion of transmembrane helices for light activation of rhodopsin” *Science* 274 (1996) 768-770.

25. Fiser A, Do RK, Sali A, “Modeling of loops in protein structures” *Prot. Sci.* 9 (2000) 1753-1773.

26. Forster MJ, “Molecular modeling in structural biology” *Micron* 33 (2002) 365-384.

27. Fujisawa Y, Kato H, Iwasaki Y, “Structure and functions of heterotrimeric G proteins in plants” *Plant Cell. Physiol.* 42:8 (2001) 789-794.

28. Garritsen A, van Galen PJ, Simonds WF, “The N-terminal coiled-coil domain of beta is essential for gamma interaction: a model for G-protein beta gamma subunit interaction” *Proc. Natl. Acad. Sci.* 90:16 (1993) 7706-7710.

29. Gomes I, Jordan BA, Gupta A, Rios C, Trapadzie N, Devi LA, “G protein coupled receptor dimerization: implications in modulating receptor function” *J. Mol. Med.* 79 (2001) 226-242.

30. Gotor C, Lam E, Cejudo FJ, Romero LC, “Isolation and analysis of the soybean SGA2 gene (cDNA), encoding a new member of the plant G-protein family of signal transducers” *Plant Mol. Biol.* 32 (1996) 1227-1234.

31. Hamm HE, “The many faces of G protein signaling” *J.Biol. Chem.* 273:2 (1998) 669-672.

32. Henikoff S, Henikoff JG, “Amino acid substitution matrices from protein blocks” *Proc. Natl. Acad. Sci. USA* 89:22 (1992) 10915-10919.

33. Hepler JR, Berman DM, Gilman AG, Kozasa T, “RGS4 and GAIP are GTPase-activating proteins for $G_{q\alpha}$ and block activation of phospholipase Cβ by γ-thio-GTP-$G_{q\alpha}$” *Proc. Natl. Acad. Sci. USA* 94 (1997) 428-432.

34. Higashijima T, Ferguson KM, Sternweis PC, Smigel MD, Gilman AG, “Effects of Mg^{2+} and the βγ-subunit complex on the interactions of guanine nucleotides with G proteins” *J. Biol. Chem.* 262 (1987) 762-766.

35. Hur EM, Kim KT, “G protein-coupled receptor signaling and cross-talk Achieving rapidity and specificity” *Cellular Signalling* 14:4 (2002) 397-405.

36. Iñiguz-Lluhi J, Kleuss C, Gilman AG, “The importance of G-protein β/γ subunits” *Trends in Cell Biology* 3 (1993) 230-235.

37. Ishikawa A, Tsubouchi H, Iwasaki Y, Asahi T, "Molecular cloning and characterization of a cDNA for the α subunit of a G protein from rice" *Plant Cell Physiol.* 36 (1995) 353-359.

38. Jones DT, "Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices" *J. Mol. Biol.* 292 (1999) 195-202.

39. Jones HD, Smith SJ, Desikan R, Plakidou-Dymock S, Lovegrove A, Hooley R, "Heterotrimeric G proteins are implicated in gibberellin induction of a-amylase gene expression in wild oat aleurone" *Plant Cell* 10 (1998) 245-253.

40. Josefsson LG, Rask L, "Cloning of a putative G-protein-coupled receptor from *Arabidopsis thaliana*" *Eur. J. Biochem.* 249:2 (1997) 415-420.

41. Karplus K, Barrett C, Hughey R, "Hidden Markov Models for Detecting Remote Protein Homologies" *Bioinformatics* 14 (1998) 846-856.

42. Katz A, Simon MI, "A segment of the C-terminal half of the G-protein β_1 subunit specifies its interaction with the γ_1 subunit" *Proc. Natl. Acad. Sci.* USA 92:6 (1995) 1998-2002.

43. Kaydamov C, Tewes A, Adler K, Manteuffel R, "Molecular characterization of cDNAs encoding G protein alpha and beta subunits and study of their temporal and spatial expression patterns in Nicotiana plumbaginifolia Viv" *Biochim. Biophys. Acta* 149 (2000) 143-160.

44. Kehrl JH, "Heterotrimeric G protein signalling: Roles in immune function and fine-tuning by RGS proteins" *Immunity* 8:1 (1998) 1-10.

45. Kikuchi A, "Modulation of Wnt signaling by Axin and Axil" *Cytokine Growth Factor Rev.* 10 (1999) 255-265.

46. Kim WY, Cheong NE, Lee DC, Je DY, Bahk JD, Cho MJ, Lee SY, “Cloning and sequencing analysis of a full-length cDNA encoding a G protein alpha subunit, SGA1, from soybean” *Plant Physiol.* 108 (1995) 1315-1316.

47. Kusnetsov VV, Oelmueller R, “Isolation of cDNAs encoding the subunit α of heterotrimeric G proteins from *Lupinus luteus* (accession no. X99485)” *Plant Physiol.* 112 (1996b) 1399.

48. Lambright DG, Noel JP, Hamm HE, Siegler PB, “Structural determinants for activation of the α subunit of a heterotrimeric G protein” *Nature* 369 (1994) 621-628.

49. Lania A, Mantovani G, Spada A, “G protein mutations in endocrine diseases” *Eur. J. Endoc.* 145 (2001) 543-559.

50. Leckie CP, McAinsh MR, Allen GJ, Sanders D, Hetherington AM, “Abscisic acid-induced stomatal closure mediated by cyclic ADP-ribose” *Proc. Natl. Acad. Sci.* USA 95 (1998) 15837-15842.

51. Lee SP, O’Dowd BF, Ng GY, Varghese G, Akil H, Mansour A, Nguyen T, George SR, “Inhibitation of cell surface expression by mutant receptors demonstrates that D_2 dopamine receptors exist as oligomers in the cell” *Mol. Pharmacol.* 58 (2000) 120-128.

52. Lee YR, Assmann SM, “Arabidopsis thaliana 'extra-large GTP-binding protein' (AtXLG1): a new class of G-protein” *Plant Mol. Biol.* 40:1 (1999) 55-64.

53. Lochrie MA, Simon MI, “G protein multiplicity in eukaryotic signal transduction systems” *Biochem.* 27 (1988) 4957-4965.

54. Ma H, Yanofsky MF, Meyerowitz EM, "Molecular cloning an characterization of GPA1, a G protein alpha subunit gene from *Arabidopsis thaliana*" *Proc. Natl. Acad. Sci. USA* 87:10 (1990) 3821-3825.

55. Ma H, Yanofsky MF, Huang H, "Isolation and sequence analysis of TGA1 cDNAs encoding a tomato G protein α subunit" *Gene* 107 (1991) 189-195.

56. MacRobbie EA, "ABA activates multiple Ca^{2+} fluxes in stomatal guard cells, triggering vacuolar $K^{+}(Rb^{+})$ release. *Proc. Natl. Acad. Sci.* USA 97 (2000) 12361-12368.

57. Malbon CC, "Heterotrimeric G-proteins and development" *Biochem. Pharmacol* 53:1 (1997) 1-4.

58. Margeta-Mitrovic M, Jan NJ, Jan LY, "Ligand-induced signal transduction within heterodimeric $GABA_B$ receptor" *Proc. Natl. Acad. Sci. USA* 98 (2001) 14643-14648.

59. Marsh JF 3rd, Kaufman LS, "Cloning and characterisation of PGA1 and PGA2: two G protein alpha-subunits from pea that promote growth in the yeast Saccharomyces cerevisiae" *Plant J.* 19 (1999) 237-247.

60. Mason MG, Botella JR, "Isolation of a novel G-protein γ-subunit from *Arabidopsis thaliana* and its interaction with Gβ" *Biochim. Biophys. Acta.* 1520 (2001) 147-153.

61. Matozaki T, Nakanashi H, Takai Y, "Small G-protein networks: Their crosstalk and signal cascades" *Cell. Signalling* 12 (2000) 515-524.

62. Mattera R, Graziano MP, Yatani A, Zhou Z, Graf R, Codina J, Birnbaumer L, Gilman AG, Brown AM, "Splice variants of the alpha subunit of the G protein Gs

activate both adenylyl cyclase and calcium channels" *Science* 243:4892 (1989) 804-807.

63. McLaughlin SK, McKinnon PJ, Margolskee RF, "Gustducin is a taste-cell-specific G protein closely related to the transducins" *Nature* 357:6379 (1992) 563-569.

64. Medkova M, Preininger AM, Yu N-J, Hubbell WL, Hamm HE, "Conformational changes in the amino-terminal helix of the G protein α_{i1} following dissociation from $G_{\beta\gamma}$ subunit and activation" *Biochemistry* 41 (2002) 9962-9972.

65. Milligan G, "Oligomerisation of G-protein-coupled receptors" *J. Cell Science* 114 (2001) 1265-1271.

66. Millner PA, "Heterotrimeric G-proteins in plant cell signalling" *New Phytol.* 151 (2001) 165-174.

67. Montell C, "PLC fills a GAP in G-protein-coupled signalling" *Nat. Cell Biol.* 2:5 (2000) E82-E83.

68. Neves SR, Ram PT, Iyengar R, "G protein pathways" *Science* 296 (2002) 1636-1639.

69. Olszewski N, Sun TP, Gubler F, "Gibberellin signaling: Biosynthesis, catabolism, and response pathways" *Plant Cell* 14 (2002) S61-S80.

70. Osguthorpe DJ, "*Ab initio* protein folding" *Curr. Opin. Struc. Biol.* 10 (2000) 146-152.

71. Pacini B, Petrigliano A, Diffley P, Paffetn A, Brown EG, Martelli P, Trabalzini L, Bovalini L, Lhusini P, Newton RP, "Adenylyl cyclase activity in *Pisum sativum*" *Phytochemistry* 34:4 (1993) 899-903.

72. Perroud PF, Crespi P, Crevecoeur M, Fink A, Tacchini P, Greppin H, "Detection and Characterization of GTP-binding proteins on tonoplast of *Spinacia oleracea*" *Plant Sci.* 122 (1997) 23-33.

73. Perroud PF, Diogon T, Crevecoeur M, Greppin H, "Molecular cloning, spatial and temporal characterization of spinach SOGA1 cDNA, encoding an α subunit of G protein"; *Gene* 248 (2000) 191-201.

74. Potenza MN, Gold SJ, Roby-Shemkowitz A, Lerner MR, Nestler EJ, "Effects of regulators of G protein-signaling proteins on the functional response of the mu-opioid receptor in a melanophore-based assay" *J. Pharmacol. Exp. Ther.* 291 (1999) 482-491.

75. Poulsen C, Mai XM, Borg S, "A *Lotus japonicus* cDNA encoding an α subunit of a heterotrimeric G-protein" *Plant Physiol.* 105 (1994) 1453-1454.

76. Rens-Domiano S, Hamm HE, "Structural and functional relationships of heterotrimeric G-proteins" *FASEB J.* 9 (1995) 1059-1066.

77. Saalbach G, Natura B, Lein W, Buschmann P, Dahse L, Rohrbeck M, Nagy F, "The α-subunit of a heterotrimeric G-protein from tobacco, NtGPα1, functions in K^+ channel regulation in mesophyll cells" *J. Exp. Bot.* 50 (1999) 53-61.

78. Sambrook J, Russell DW, "Molecular Cloning: A laboratory manual" *Cold Spring Harbor Laboratory Press NY* 3rd ed. (2001).

79. Sanchez R, Sali A, "Advances in comparative protein-structure modeling" *Curr. Opin. Struc. Biol.* 7 (1997) 206-214.

80. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF, "Improving the accuracy of PSI-BLAST protein database

searches with composition-based statistics and other refinements" *Nucl. Acids Res.* 29:14 (2001) 2994-3005.

81. Schulz A, Grosse R, Schultz G, Gudermann T, Schoneberg T, "Structural implication for receptor oligomerization from functional reconstitution studies of mutant V_2 vasopressin receptors" *J. Biol. Chem.* 275 (2000) 2381-2389.

82. Seo HS, Kim HY, Jeong JY, Lee SY, Cho MJ, Bahk JD, "Molecular cloning and characterization of RGA1 encoding a G protein alpha subunit from rice (Oryza sativa L. IR-36)" *Plant Mol. Biol.* 27 (1995) 1119-1131.

83. Smrcka AV, Stemweis PC, "Regulation of purified subtypes of phosphatidylinositol-specific phospholipase C β by G protein α and βγ subunits" *J. Biol. Chem.* 268:13 (1993) 9667-9674.

84. Sondek J, Lambright DG, Noel JP, Hamm HE, Sigler PB, "GTPase mechanism of G proteins from the 1.7-A crystal structure of transducin alpha-GDP-AIF-4" *Nature* 372:6503 (1994) 276-279.

85. Sondek J, Bohm A, Lambright DG, Hamm HE, Sigler PB, "Crystal structure of a G-protein beta gamma dimer at 2.1 A resolution" *Nature* 379:6563 (1996) 369-374.

86. Sprang SR, "G protein mechanisms: Insights from structural analysis" *Annu. Rev. Biochem.* 66 (1997) 639-678.

87. Stern KR, "Introductory plant biology" *Wm. C. brown Publishers* 7th ed. (1997) 182-184.

88. Tang WJ, Gilman AG, "Adenylyl cyclases" *Cell* 70 (1992) 869-872.

89. Thompson JD, Higgins DG, Gibson TG, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice" *Nucl. Acids Res.* 22:22 (1994) 4673-4680.

90. Topham CM, Thomas P, Overington JP, Johnson MS, Eisenmenger F, Blundell TL, "An assessment of COMPOSER: a rule-based approach to modeling protein structure" *J. Molec. Biol.* 229 (1990) 194-220.

91. Venclovas C, Zemla A, Fidelis K, Moult J, "Some measures of comparative performance in the three CASPs" *Proteins* Suppl. 3 (1999) 231-237.

92. Wall MA, Coleman DE, Lee E, Iñiguz-Lluhi JA, Posner BA, Gilman AG, Sprang SR, "The structure of the G protein heterotrimer $G_{i\alpha 1}\beta_1\gamma_2$" *Cell* 83 (1995) 1047-1058.

93. Wang X-Q, Wu W-H, Assmann SM, "Differential responses of abaxial and adaxial guard cells of broad bean to abscisic acid and calcium" *Plant Physiol.* 118 (1998) 1421-1429.

94. Wang X-Q, Ullah H, Jones AM, Assmann SM, "G protein regulation of ion channels and abscisic acid signaling in *Arabidopsis* guard cells" *Science* 292 (2001) 2070-2072.

95. Whiteway M, Clark KL, Leberer E, Dignard D, Thomas DY, "Genetic identification of residues involved in association of α and β G-protein subunits" *Mol. Cell. Biol.* 14:5 (1994) 3223-3229.

96. Wickman K, Clapham DE, "Ion channel regulation by G proteins" *Physiol. Rev.* 75:4 (1995) 865-885.

97. Wilkie TM, “G-protein signaling: Satisfying the basic necessities of life” *Current Biology* 10 (2000) R853-R856.

98. Wylie F, Heimann K, Le TL, Brown D, Rabnott G, Stow JL, “GAIP, a $G\alpha_{i-3}$-binding protein, is associated with Golgi-derived vesicles and protein trafficking” *Am. J. Physiol.* 276 (1999) C497-C506.

99. Zhong H, Neubig RR, “Regulator of G protein signaling proteins: Novel Multifunctional Drug Targets” *J. Pharmacol. Exp. Ther.* 297:3 (2001) 837-845.

100. Zimmermann H, “Synaptic transmission: Cellular and molecular basis” 1st ed. *Oxford University Press* Jan.1993.

APPENDIX A

PSIBLAST iteratively searches one or more protein databases for sequences similar to one or more protein query sequences. PSIBLAST is similar to BLAST except that it uses position-specific scoring matrices derived during the search. PSIBLAST uses position-specific scoring matrices (PSSMs) to score matches between query and database sequences, in contrast to BLAST that uses pre-defined scoring matrices such as BLOSUM62. PSIBLAST may be more sensitive than BLAST, meaning that it might be able to find distantly related sequences that are missed in a BLAST search. (Altschul SF *et al.*, 1997)

PSIBLAST can repeatedly search the target databases, using a multiple alignment of high scoring sequences found in each search round to generate a new PSSM for use in the next round of searching. PSIBLAST will iterate until no new sequences are found, or the user specified maximum number of iterations is reached, whichever comes first. Normally, the first round of searching uses a standard scoring matrix, effectively performing a blast search.

PSIBLAST is a statistically driven search method that finds regions of similarity between a given query sequence and database sequences and produces gapped alignments of those regions. Within these aligned regions, the calculated score is higher than some level that one would expect to occur by chance alone. (Schaffer AA *et al.*, 2001)

A typical PSIBLAST search starts with the selection of the sequence of interest, and the database(s) in which the search is going to be performed. There are different databases available and new ones are being added continuously. The most common databases are the PDB Sequences, GenPept Full Release, TrEMBL, GenBank Bacterial,

Viral, Other Mammalian, Rodent, Other Invertebrate Sequences and Genome Databases. The search itself requires several parameters to be set, of which the most important ones are:

Maximum expectation level (This value is set for each round and defines the expectation of a sequence as the probability of the current search finding a sequence with a score that is obtained by chance alone. Setting the maximum expectation level to 10.0, the default, limits the reported sequences to those with scores high enough to have been found by chance only ten or fewer times)

Threshold expectation value (A value that is used for obtaining sequences to score and still be used to build PSSMs. Typically, this threshold is a smaller value than the maximum expectation level and the default is 0.005)

Maximum number of rounds (Specifies the repetition number of searches within the selected databases. Default value is 5)

Matrix selection (After each search round, high-scoring sequences are used to create a multiple alignment that is then used to calculate match scores for the PSSM. When building the PSSMs, part of each score is based upon observed amino acid frequencies in the multiple alignments, and part is based on prior knowledge of amino acid substitutability. The prior information, represented as "pseudo counts", is derived from a standard scoring matrix, such as BLOSUM62. (Henikoff S and Henikoff JG, 1992) Pseudo counts are particularly useful when the sequences included in the multiple alignments do not constitute an adequate sample of the protein family that they represent. Other available matrices are BLOSUM 45, BLOSUM 80, PAM 40 and PAM 70)

The parameters and the database used in this study is given below, together with the PSIBLAST output.

Database: PDB Sequences

Max. Expectation Value: 10 (default)

Threshold expectation value: 0.005 (default)

Number of rounds: 5 (default)

Matrix: BLOSUM 62 (default)

Iteration	Used in model and found again	Hits not used in model
1		52
2	38	18
3	48	28
4	52	48

Database	ID	Name	Score	E-value
PDBSEQRES	1GG2_A	COMPLEX (GTP-BINDING/TRANSDUCER)	554	e-158
PDBSEQRES	1FQK_A	SIGNALING PROTEIN	554	e-158
PDBSEQRES	1AGR_A	COMPLEX (SIGNAL TRANSDUCTION/REGULATOR)	554	e-158
PDBSEQRES	1CIP_A	HYDROLASE	554	e-158
PDBSEQRES	1GDD__	SIGNAL TRANSDUCTION PROTEIN	554	e-158
PDBSEQRES	1GFI__	SIGNAL TRANSDUCTION PROTEIN	554	e-158
PDBSEQRES	1GIA__	SIGNAL TRANSDUCTION PROTEIN	554	e-158
PDBSEQRES	1GP2_A	COMPLEX (GTP-BINDING/TRANSDUCER)	554	e-158
PDBSEQRES	1GIL__	GTP-BINDING PROTEIN	551	e-157
PDBSEQRES	1BOF__	SIGNAL TRANSDUCTION PROTEIN	551	e-157
PDBSEQRES	1GIT__	GTP-BINDING PROTEIN	551	e-157
PDBSEQRES	1AS0__	SIGNAL TRANSDUCTION	550	e-157
PDBSEQRES	1AS2__	SIGNAL TRANSDUCTION	550	e-157
PDBSEQRES	1AS3__	SIGNAL TRANSDUCTION	550	e-157

PDBSEQRES	1FQJ_A	SIGNALING PROTEIN	531	e-151
PDBSEQRES	1FQJ_D	SIGNALING PROTEIN	531	e-151
PDBSEQRES	1AGR_D	COMPLEX (SIGNAL TRANSDUCTION/REGULATOR)	531	e-151
PDBSEQRES	1FQK_C	SIGNALING PROTEIN	531	e-151
PDBSEQRES	1KJY_A	SIGNALING PROTEIN	529	e-151
PDBSEQRES	1KJY_C	SIGNALING PROTEIN	529	e-151
PDBSEQRES	1GOT_A	COMPLEX (GTP-BINDING/TRANSDUCER)	519	e-148
PDBSEQRES	1TAD_A	GTP-BINDING PROTEIN	515	e-147
PDBSEQRES	1TAD_B	GTP-BINDING PROTEIN	515	e-147
PDBSEQRES	1TAD_C	GTP-BINDING PROTEIN	515	e-147
PDBSEQRES	1TAG__	GTP-BINDING PROTEIN	515	e-147
PDBSEQRES	1TND_A	BINDING PROTEIN(GTP)	515	e-147
PDBSEQRES	1TND_B	BINDING PROTEIN(GTP)	515	e-147
PDBSEQRES	1TND_C	BINDING PROTEIN(GTP)	515	e-147
PDBSEQRES	1BH2__	SIGNAL TRANSDUCTION PROTEIN	509	e-145
PDBSEQRES	1AZS_C	COMPLEX (LYASE/HYDROLASE)	473	e-134
PDBSEQRES	1AZT_A	HYDROLASE	473	e-134
PDBSEQRES	1AZT_B	HYDROLASE	473	e-134
PDBSEQRES	1CUL_C	LYASE/LYASE/SIGNALING PROTEIN	472	e-134
PDBSEQRES	1CJK_C	LYASE/LYASE/SIGNALING PROTEIN	468	e-132
PDBSEQRES	1CJT_C	LYASE/LYASE/SIGNALING PROTEIN	468	e-132
PDBSEQRES	1CJU_C	LYASE/LYASE/SIGNALING PROTEIN	468	e-132
PDBSEQRES	1CJV_C	LYASE/LYASE/SIGNALING PROTEIN	468	e-132
PDBSEQRES	1CS4_C	LYASE/LYASE/SIGNALING PROTEIN	464	e-131
PDBSEQRES	1HFV_A	G PROTEIN	133	1e-31
PDBSEQRES	1HFV_B	G PROTEIN	133	1e-31
PDBSEQRES	1E0S_A	G PROTEIN	133	1e-31
PDBSEQRES	1RRF__	TRANSPORT PROTEIN	130	9e-31

PDBSEQRES	1RRG_A	TRANSPORT PROTEIN	130	9e-31
PDBSEQRES	1RRG_B	TRANSPORT PROTEIN	130	9e-31
PDBSEQRES	1HUR_A	PROTEIN TRANSPORT	129	9e-31
PDBSEQRES	1HUR_B	PROTEIN TRANSPORT	129	9e-31
PDBSEQRES	1KSG_A	SIGNALING PROTEIN/HYDROLASE	121	4e-28
PDBSEQRES	1KSH_A	SIGNALING PROTEIN/HYDROLASE	121	5e-28
PDBSEQRES	1KSJ_A	SIGNALING PROTEIN/HYDROLASE	118	4e-27
PDBSEQRES	1F6B_A	PROTEIN TRANSPORT	112	2e-25
PDBSEQRES	1F6B_B	PROTEIN TRANSPORT	112	2e-25
PDBSEQRES	1FZQ_A	SIGNALING PROTEIN	107	7e-24

Evaluation of the output:

Bit Score (Score): Each aligned segment pair has a normalized score expressed in bits, which lets one estimate the magnitude of the search space before expecting to find an HSP score as good as or better than this one by chance. Thus, higher bit score results in better matches.

E-value: There is a probability associated with each pair wise comparison in the list and with each segment pair alignment. The number shown in the list is the probability of observing a score purely by chance when a search is performed against a database of this size. Higher E-values, approaching 1.0, indicate a score that would be obtained by chance alone.

CLUSTALW performs multiple sequence alignment on protein or nucleic sequences.

General Parameters

Output Order

Used to control the order of the sequences in the output alignments. By default, the order corresponds to the order in which the sequences were aligned (from the guide tree/dendrogram), thus automatically grouping closely related sequences. This switch can be used to set the order to the same as the input file.

Pairwise Alignment Parameters

Alignment Method

The choices are an accurate, dynamic programming method, or the faster, approximate method of Wilbur and Lipman. The dynamic method can be slow for a large number of long sequences, e.g. 20 genetic sequences of over 1000 residues.

Accurate Parameters

These parameters do not have any affect on the speed of the alignments. They are used to give initial alignments that are then rescored to give percent identity scores. These percentage scores are the ones, which are displayed on the screen. The scores are converted to distances for the trees.

Gap Open Penalty

The penalty for opening a gap in the alignment.

Gap Extension Penalty

The penalty for extending a gap by 1 residue.

Protein Weight Matrix

The scoring table which describes the similarity of each amino acid to each other.

DNA Weight Matrix

The scores assigned to matches and mismatches (including IUB ambiguity codes).

Fast parameters

These similarity scores are calculated from fast, approximate, global alignments, which are controlled by 4 parameters. Two techniques are used to make these alignments very fast: 1) only exactly matching fragments (k-tuples) are considered 2) only the 'best' diagonals (the ones with most k-tuple matches) are used.

K-tuple Size

This is the size of exactly matching fragment that is used. INCREASE for speed (max= 2 for proteins; 4 for DNA), DECREASE for sensitivity. For longer sequences (e.g. >1000 residues) you may need to increase the default.

Gap Penalty

This is a penalty for each gap in the fast alignments. It has little affect on the speed or sensitivity except for extreme values.

Top Diagonals

The number of k-tuple matches on each diagonal (in an imaginary dot-matrix plot) is calculated. Only the best ones (with most matches) are used in the alignment. This parameter specifies how many. Decrease for speed; increase for sensitivity.

Window Size

This is the number of diagonals around each of the 'best' diagonals that will be used. Decrease for speed; increase for sensitivity.

Multiple Alignment Parameters

These parameters control the final multiple alignment. This is the core of the program and the details are complicated. To understand the use of the parameters and the scoring system, you will have to refer to the additional documentation.

Each step in the final multiple alignment consists of aligning two alignments or sequences. This is done progressively, following the branching order in the GUIDE TREE. The basic parameters to control this are two gap penalties and the scores for various identical / non-identical residues.

Gap Open Penalty

The penalty for opening a gap in the alignment. Increasing the gap-opening penalty will make gaps less frequent.

Gap Extension Penalty

The penalty for extending a gap by 1 residue. Increasing the gap extension penalty will make gaps shorter. Terminal gaps are not penalized.

Delay Divergent Sequences

Delays the alignment of the most distantly related sequences until after the most closely related sequences have been aligned. The setting shows the percent identity level required to delay the addition of a sequence; sequences that are less identical than this level to any other sequences will be aligned later.

DNA Transition Weight

Gives transitions (A <--> G or C <--> T i.e. purine-purine or pyrimidine-pyrimidine substitutions) a weight between 0 and 1; a weight of zero means that the transitions are scored as mismatches, while a weight of 1 gives the transitions the match score. For distantly related DNA sequences, the weight should be near to zero; for closely related sequences, it can be useful to assign a higher score.

Protein Gap Parameters

Gap Penalty options which are only used in protein alignments.

Residue Specific Penalties are amino acid specific gap penalties that reduce or increase the gap opening penalties at each position in the alignment or sequence. See the additional documentation for details. As an example, positions that are rich in glycine are more likely to have an adjacent gap than positions that are rich in valine.

Hydrophilic Gap Penalties are used to increase the chances of a gap within a run (5 or more residues) of hydrophilic amino acids; these are likely to be loop or random coil regions where gaps are more common. The residues that are "considered" to be hydrophilic are set by menu item 3.

Gap Separation Distance tries to decrease the chances of gaps being too close to each other. Gaps that are less than this distance apart are penalised more than other gaps. This does not prevent close gaps; it makes them less frequent, promoting a block-like appearance of the alignment.

End Gap Separation treats end gaps just like internal gaps for the purposes of avoiding gaps that are too close (set by GAP SEPARATION DISTANCE above). If you turn this off, end gaps will be ignored for this purpose. This is useful when you wish to align fragments where the end gaps are not biologically meaningful.

Weight Matrices

For protein alignments, a weight matrix is used to determine the similarity of non-identical amino acids. For example, Tyr aligned with Phe is usually judged to be 'better' than Tyr aligned with Pro.

There are three 'in-built' series of weight matrices offered. Each consists of several matrices, which work differently at different evolutionary distances. Crudely, several matrices are stored in memory, spanning the full range of amino acid distance (from almost identical sequences to highly divergent ones). For very similar sequences, it is best to use a strict weight matrix that only gives a high score to identities and the most favored conservative substitutions. For more divergent sequences, it is appropriate to use "softer" matrices, which give a high score to many other frequent substitutions.

Blosum

These matrices appear to be the best available for carrying out data base similarity (homology searches). The matrices used are the Blosum80, 62, 45 and 30 matrices.

PAM

These have been extremely widely used since the late '70s. The matrices used are the PAM 120, 160, 250 and 350 matrices.

Gonnet

These matrices appear to be more sensitive than the PAM series. The matrices used are the Gonnet 40, 80, 120, 160, 250 and 350 matrices.

Identity

This matrix gives a score of 1.0 to two identical amino acids and a score of zero otherwise. This matrix is not very useful.

For DNA, a single matrix (not a series) is used. Two hard-coded matrices are available.

IUB

This is the default scoring matrix used by BESTFIT for the comparison of nucleic acid sequences. X's and N's are treated as matches to any IUB ambiguity symbol. All matches score 1.9; all mismatches for IUB symbols score 0.

The parameters used in this study are:

Weight matrix: Blosum series
Gap open penalty: 10.00
Gap extension penalty: 0.20
Delay divergent sequences: 30
Residue-specific gap penalties: on
Hydrophilic gap penalties: on
End gap separation penalty: off

APPENDIX B

Autoclave:	Hirayama, Hiclave HV-110, JAPAN
	Certoclav, Table Top Autoclave CV-EL-12L, AUSTRIA
Balance:	Sartorius, BP211D, GERMANY
	Sartorius, BP221S, GERMANY
	Sartorius, BP610, GERMANY
	Schimadzu, Libror EB-3200 HU, JAPAN
Centrifuge:	Eppendorf, 5415C, GERMANY
	Eppendorf, 5415D, GERMANY
	Eppendorf, 5415R, GERMANY
	Kendro Lab. Prod., Heraeus Multifuge 3L, GERMANY
	Hitachi, Sorvall RC5C Plus, USA
	Hitachi, Sorvall Discovery 100 SE, USA
Deepfreeze:	-70^{o} C, Kendro Lab. Prod., Heraeus Hfu486 Basic, GERMANY

	-20^{o} C, Bosch, TÜRKİYE
Distilled Water:	Millipore, Elix-S, FRANCE
	Millipore, MilliQ Academic, FRANCE
Electrophoresis:	Biogen Inc., USA
	Biorad Inc., USA
Gel Documentation:	UVITEC, UVIdoc Gel Documentation System, UK
	Biorad, UV-Transilluminator 2000, USA
Ice Machine:	Scotsman Inc., AF20, USA
Incubator:	Memmert, Modell 300, GERMANY
	Memmert, Modell 600, GERMANY
Laminar Flow:	Kendro Lab. Prod., Heraeus, HeraSafe HS12, GERMANY
Magnetic Stirrer:	VELP Scientifica, ARE Heating Magnetic Stirrer, ITALY
	VELP Scientifica, Microstirrer, ITALY
Microliter Pipette:	Gilson, Pipetman, FRANCE
	Mettler Toledo, Volumate, USA
Microwave Oven:	Bosch, TÜRKİYE

pH meter:	WTW, pH540 GLP MultiCal®, GERMANY
Power Supply:	Biorad, PowerPac 300, USA
	Wealtec, Elite 300, USA
Refrigerator:	+4° C, Bosch, TÜRKİYE
Shaker:	Forma Scientific, Orbital Shaker 4520, USA
	GFL, Shaker 3011, USA
	New Brunswick Sci., Innova™ 4330, USA
Spectrophotometer:	Schimadzu, UV-1208, JAPAN
	Schimadzu, UV-3150, JAPAN
	Secoman, Anthelie Advanced, ITALY
Speed Vacuum:	Savant, Speed Vac® Plus Sc100A, USA
	Savant, Refrigerated Vapor Trap RVT 400, USA
Thermocycler:	Eppendorf, Mastercycler Gradient, GERMANY
Vacuum:	Heto, MasterJet Sue 300Q, DENMARK
Water bath:	Huber, Polystat cc1, GERMANY