Biochemistry

 Volume 30, Number 19 May 14, 1991

Perspectives in Biochemistry

The *ras* Protein Family: Evolutionary Tree and Role of Conserved Amino Acids

Alfonso Valencia,[‡] Pierre Chardin,[§] Alfred Wittinghofer,[∥] and Chris Sander*[,‡]

European Molecular Biology Laboratory, Meyerhofstrasse 1, D-6900 Heidelberg, Germany, Institut de Pharmacologie du CNRS, route des Lucioles, Sophia-Antipolis, F-06560 Valbonne, France, and Max Planck Institute für medizinische Forschung, Abteilung Biophysik, Jahnstrasse 29, D-6900 Heidelberg, Germany

Received January 16, 1991

A large number of different GTP/GDP-binding proteins is present in eukaryotic cells. These G-proteins are members of at least five distinct families: the elongation factors of protein biosynthesis, subunits of the signal recognition particle (SRP) and its receptor, the ADP-ribosylation factor (ARF) family, the α subunits of heterotrimer G-proteins directly involved in signal transduction, and the products of the *ras* gene family.

The early discovery of *ras* genes was due to their highly oncogenic potential when transduced in retroviruses, such as the Harvey and Kirsten murine sarcoma viruses containing viral forms of H-*ras* and K-*ras*, respectively. Variants of these two genes and of a third one, N-*ras*, not previously found in a retrovirus, were characterized independently as the transforming genes present in many human or animal tumors. In 10%–50% of human tumors, one of the three endogenous *ras* genes, H-*ras*, K-*ras*, or N-*ras*, is activated by a somatic point mutation leading to the substitution of a single amino acid, usually in position 12 or 61, and plays an important role in the acquisition of a transformed phenotype (Barbacid, 1987; Bos et al., 1988; Chardin, 1988; Spandidos, 1989). Four main approaches have led to the isolation of new members of the *ras* family: serendipity, homology probing, protein purification, and characterization of yeast mutants. With these approaches the primary structures of more than 50 *ras*-related proteins are now available from cloned cDNA sequences.

The three-dimensional structure of the G-domain (the guanine nucleotide binding) of the H-*ras* p21 protein bound to various guanine nucleotides [residues 1–171, Milburn et al. (1990); residues 1–166, Pai et al. (1989, 1990) and Schlichting et al. (1990)] has now been determined. The N- and C-terminal parts of the sequence form two subdomains of the structure (front and back in Figure 1). The guanine nucleotide spans almost the entire width of the structure and is in contact with both subdomains. The interactions between protein and guanine nucleotide have been defined in great detail and the mechanism of GTP hydrolysis has been proposed on the basis of the high-resolution structure (Pai et al., 1990). Comparison of the three-dimensional structures of p21 and of the G-domain of bacterial elongation factor Tu (EF-Tu; LaCour et al., 1985; Jurnak, 1985) reveals that the two G-domains share a conserved topology and that 65% of the residues are structurally equivalent (Valencia et al., 1991). This makes it very likely that the G-binding domains of all G-proteins have the same basic structure.

It therefore is timely to compare the primary sequences of the members of the *ras* gene family, to construct a phylogenetic tree, and to describe the role of the conserved residues. We define conserved and variable regions and show where these are located in the three-dimensional structure of the *ras* protein p21 (deVos et al., 1988; Pai et al., 1989, 1990; Milburn et al., 1990; Schlichting et al., 1990). This should facilitate detailed understanding of the functional role of particular residues and planning of mutational experiments. In addition, the alignment of primary structures allows us to classify the small guanine nucleotide binding proteins into four main branches and various subbranches and to predict functional specificities for each branch of this family.

Main Sequence Regions

The primary sequence of ras related proteins may be subdivided into several main parts (Figure 2; all residue numbers, unless otherwise stated, refer to H-*ras* p21). The N-terminal end, before the first highly conserved residue, Lys5, is of variable length ranging from these residues to more than 30 (Figure 3B). In the three-dimensional structure these N-terminal extensions can form protrusions just before β strand β1 and may be involved in interactions with other proteins. The N-terminal subdomain has the first foru β strands and two helices, from residue K5 to I84, and contains the two phosphate/Mg-binding loops with the conserved boxes GxxxxGK[S,T] and DTAG (Figure 3A). The C-terminal

[‡] EMBL.
[§] Institut de Pharmacologie du CNRS.
[∥] Max Planck Institute für medizinische Forschung.

0006-2960/91/0430-4637$02.50/0

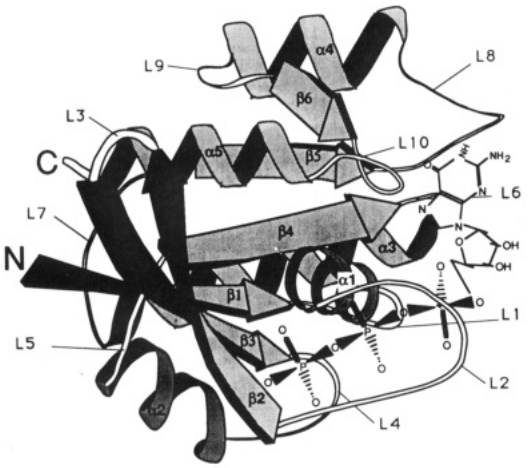

FIGURE 1: Cartoon representation of the structure of *ras* p21 protein, residues 1–166, as determined by Pai et al. (1990). Secondary structure elements $\alpha 1$–$\alpha 5$ are helices, $\beta 1$–$\beta 6$ are β strands, and L1–L10 are loop regions. The corresponding stretches of amino acid sequence are identified by the same labels in Figure 3. The most conserved sequences boxes are in loops L1, L4, L6, and L10, on the right side of the cartoon. The position and conformation of the bound GTP nucleotide is sketched approximately. Adapted from a drawing by Doug Lowy.

subdomain has the remaining two β strands and two helices, from residue N85 to about R164, and contains the guanine base binding loops NKxD and ExSAK (Figure 3A). The C-terminal extension, from about residue 165 to the C-terminus, is highly variable, even among closely related proteins such as H-*ras*, K-*ras*, and N-*ras*, ranging in length from 14 to more than 130 residues. The last four residues always include a cysteine motif that appears to be required for in vivo covalent modification (Figure 3B).

The C-terminus of p21 is farnesylated, clipped, carboxymethylated, and palmitoylated. These modifications apparently are required for anchoring the C-terminus in the membrane (Gutierrez et al., 1989; Hancock et al., 1989; Schafer et al., 1989). The full-length protein containing the C-terminus has been crystallized (Milburn et al., 1990). However, the structure of the C-terminal region is partially disordered, suggesting that it is mobile, sticks out of the globular G-binding domain, and may act as a flexible spacer between the membrane-bound C-terminus and the globular domain in the cytoplasm. The complete C-terminal region is believed to provide the specific signal for posttranslational modification and for the interaction of each *ras*-like protein with its proper membrane. The biochemical properties of the G-binding domain of *ras* p21 are independent of the presence of the C-terminus (John et al., 1988).

Multiple Sequence Alignment of G-Domains

All available sequences in the *ras* family were aligned by using a dynamic programming algorithm (Sander & Schneider, 1991; Smith & Waterman, 1981). This algorithm seeks to optimize amino acid similarities, not only identities, summed over all aligned sequence positions, for two sequences at a time. As the basic three-dimensional structure is likely to be conserved in the entire family, no insertions and deletions are permitted in the known secondary structure elements. The very variable N- and C-terminal tails cannot be meaningfully aligned but are listed for completeness.

The alignment of the sequences of 32 selected *ras*-related proteins is shown in Figure 3, together with the secondary structural elements and residue solvent accessibility of *ras* p21. The solvent accessibility number gives a rough estimate of the number of water molecules in contact with a residue. Sequence motifs involved in the binding of Mg^{2+} and phosphate are labeled PM1, PM2, and PM3, and those involved in binding of the guanine base, G1 and G2. The presence of these sequence regions can be used to identify a protein sequence as belonging to the general class of G-domains.

The variability of each sequence position is calculated from the multiple alignment (Figure 3) and mapped onto the structure in Figure 4. It takes into account conservative replacements of amino acids and is reported here on a scale of 0–5 (Sander & Schneider, 1991). There are 22 very strongly conserved sequence positions (variability score of zero). Of these, seven are directly involved in the interaction with nucleotide, namely, K16, F28, T35, D57, G60, K117, and D119. The variability profile also makes apparent regions of higher sequence diversity, e.g., the region including the end of helix $\alpha 1$, the loop L2, and the beginning of strand $\beta 2$. L2 is the "effector loop" believed to interact with effector proteins like GAP. Each type of G-protein presumably interacts specifically with its own GAP-like protein (Trahey and McCormick, 1987, Kikuchi et al., 1989, Garret et al. 1989, Hall 1990). The known sequences in the *ypt/rab* branch have particularly high diversity in the effector loop, indicating a wide range of specificities.

Construction of a Phylogenetic Tree

Because of the diversity of functions, species origins, and tissue specificity in the *ras* protein family, the precise time-ordered evolutionary tree is unknown and very difficult to derive. However, on the basis of the multiple sequence alignment, one can derive a family tree. Closeness in the tree can be interpreted in terms of similarity of function and/or in terms of similarity of species or cell lineage (Figure 5).

The input to the three algorithm is a set of distances. Here, the distance between any two sequences is taken to be the number of amino acid mismatches (unequal residues) in the aligned region (residues 5–164). The mathematical problem of tree construction is that of finding a tree that best represents the given interprotein distances. The problem can be solved approximately by the methods of maximum parsimony or maximum likelihood (Fitch & Margoliash, 1967; Felsenstein, 1981). Several trees were derived, for all sequences as well as for selected subsets. The results differ slightly, but the

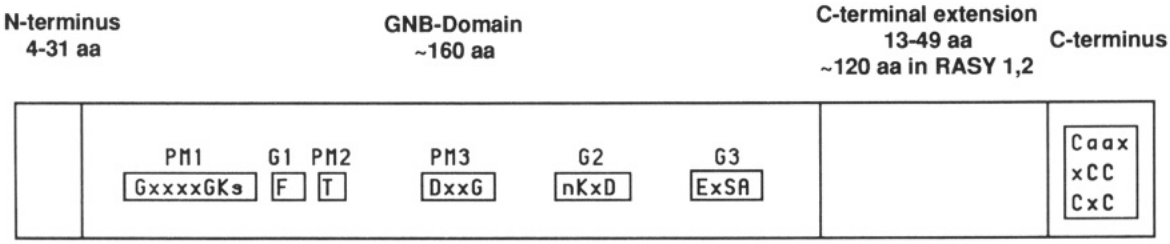

FIGURE 2: Main sequence elements and very conserved residues of *ras*-like proteins. GNB = guanine nucleotide binding domain; PMn = phosphate/magnesium binding regions; Gn = guanine base binding regions. Caax: a = aliphatic, x = any residue. GxxxGKs: s = Ser or Thr.

overall picture remains stable. Consistency of the subfamilies implied by the tree is evident in that almost all insertions and deletions are common to a subfamily, e.g., the 9–12-residue insert in the *rho* subfamily after residue 122.

From what we know about the function of *ras*-like proteins, the tree primarily represents functional, not species, relationships. For example, yeast *ypt1*, maize *yptm*, and mammalian *rab1*, although from widely different species, are quite close, indicating near-identity of cellular function, probably involvement in interorganelle transport or secretion (Hall, 1990). On the other hand, yeast *ras1* and yeast *ypt1*, although from the same species, are quite distant. We conclude that in this protein family similarity of function and not of species is the main criterion for sequence similarity.

The tree can be interpreted in terms of subgroups, or clusters, of sequences, at a particular level of detail. For example, at the level of 90% sequence identity, H-*ras*, H-*ras C*, and K-*ras* form a subgroup; other subgroups at this level are *rap1A* and *rap1B*, *rac1* and *rac2*, *rho1A* and *rho1C*, *rab1A* and *rab1B*, and *sas1* and *sas2*. Proteins in such subgroups are likely to have very similar function. At the other extreme, at 30% sequence identity, four major subgroups can be discerned: *ras*, *rho*, *ypt*, and *tc4*. Members of each of these wider subgroups probably share some general properties, while their functions may differ in detail.

rab7 is grouped with the *ypt* subfamily at a very distant level and therefore could be in a separate functional class. A recent sibling of the superfamily, *tc4*, from a human teratocarcinoma cell line (Drivas et al., 1990), is as distant from all three subfamilies as they are among each other and has a distinctive effector loop sequence (FEKKYVAT), so *tc4* almost certainly represents a fourth functional subfamily.

Number of *ras* Family Proteins in Mammals

The number of sequenced *ras*-related proteins has steadily risen from 1985 to 1990 (see Figure 5 caption). A total of more than 30 different *ras*-like genes has already been found in mammalian cells, and many more may be discovered. However, different groups working independently have now on a number of occasions isolated the same protein by different approaches; e.g., a number of small G-proteins isolated biochemically turned out to be identical with those predicted from cDNAs isolated by homology probing.

A particularly striking example of triple discovery is the *rap1A* gene that was first isolated by low-stringency hybridization with a D-*ras3* probe (Pizon et al., 1988b), then as a cDNA able to revert v-K-*ras* transformation, K-*rev1* (Kitayama et al., 1989), and as *smg* p21 (Kawata et al., 1988), a small G-protein found in brain, neutrophils and platelets. The fact that isolation of small G-proteins, by various approaches and from diverse cell types, leads to the rediscovery of already known proteins suggests that a significant fraction of this family may already have been discovered.

A *ras* Ancestral Gene in *Escherichia coli*?

Many ras and ras related proteins of mammals have closely related homologs in *Drosophila*, yeast, and *Dictyostelium*. Apparently, this type of protein already existed before the divergence of the phyla leading to insects and vertebrates, plants, yeasts, and slime molds. Does a putative common ancestor for the *ras* family exist in a more primitive organism? We have used the oligonucleotide strategy described by Touchot et al. (1987) to search for *ras*-related genes encoding a protein with a DTAGQE sequence in *E. coli*.

Under strigency conditions where most *ras*-related cDNAs were detected by an oligonucleotide mix, we failed to detect strongly hybridizing sequences in *E. coli* DNA. By lowering the strigency further, at least 5 different bands were revealed, with several restriction enzymes. The most suitable restriction enzymes were chosen for each of them, and the bands were electroeluted and cloned in plasmid or phage λ, subcloned in M13, and sequenced. None of the hybridizing regions perfectly matched the oligonucleotide sequence, and thus none encoded a perfect DTAGQE motif. Some of the clones encoded part of a DTAGQE sequence but not in an open reading frame or in an open reading frame that had significant sequence similarity to *ras* proteins (P. Chardin, unpublished).

These results suggest that there are no *ras*-like proteins in *E. coli* with a DTAGQE sequence but do not rule out the possibility that other *ras*-related proteins exist in *E. coli*. However, other *ras* probes (H-*ras*, K-*ras*, and N-*ras*) failed to detect major hybridizing sequences in *E. coli* DNA, even at very low stringency. A strongly hybridizing band was detected with the *rhoA* probe, but again sequencing failed to reveal any open reading frame with homology to *rho* proteins.

Two proteins with sequences characteristic of a GTP/GDP-binding site, in addition to the ribosomal factors, have been found in *E. coli*: the Era (Ahnn et al. 1986) and LepA (March and Inouye, 1985) proteins. However, both proteins have little similarity to the *ras* family outside of the three main nucleotide binding motifs, and they are no more closely related to *ras* than to other G-domains such as ARF, Sar1p, or the α subunits of heterotrimeric G-proteins (Figure 6). The LepA protein may be involved in protein export, as the gene is cotranscribed with a signal peptidase, but neither LepA nor Era is likely to have a function analogous to that of *ras*, *rab*, or *rho* in mammalian or yeast cells. We conclude that no close homologue of *ras* exists in *E. coli*.

Conserved Regions Involved in G Nucleotide Binding

The highest degree of sequence conservation is found in four regions that are directly involved in guanine nucleotide binding (Figures 4, 6, and 7A). The first two constitute most of the phosphate and Mg^{2+} binding site (PM site) and are located in the first half of the G-domain. The other two regions are involved in guanosine binding and are located in the C-terminal half of the molecule.

The GxxxxGK[S,T] motif, where [S,T] means S or T in this position, is found in all *ras*-related proteins and in G proteins of other function, such as ARFs, Sar1p, and Gα subunits, as well as in other nucleotide-binding proteins, such as ATPases and kinases (Gay & Walker, 1983; Wierenga et al., 1986; Dreusicke & Schulz, 1986). In *ras* p21, this region, the PM1 site, adopts a stable loop structure with the side chain of Lys 16 noncovalently closing the loop by interacting with the main-chain carbonyl groups of Gly10 and Ala11 (Pai et al., 1989). Lys16 also is in contact with the β,γ-phosphate oxygens and is presumably involved in catalysis (Pai et al., 1990; Reinstein et al., 1990). In the *ras* subfamily, most proteins have Gly–Gly in position 12–13. In *ras* p21, replacement of Gly12 by any other amino acid, except Pro, leads to a transforming potential (Seeburg et al., 1984).

What effect do these oncogenic mutations have on the three-dimensional structure? Is there a major rearrangement of the PM site? Since in the wild-type (c-H-ras) structure the backbone angles of Gly12 ($\phi = -60°$, $\varphi = +132°$) have values allowed for any amino acid, no major rearrangement is necessary, and indeed, in the crystal structures of two Gly12 mutants (Arg12 and Val12) the structure of the loop is not significantly disturbed (Krengel et al., 1990). The main effect of side chains at position 12 apparently is to physically block

A

```
         β1    L1          α1         L2              β2     L3        β3          L4      α2        L5   β4    L6
struc    EEEEE          HHHHHHHHH  H              EEEEEEE  EEE   --EEE  EEEEEEE          HHHHHH  HHH   EEEEE  EE     HHHHH
water    4000001460 2132002123 **984****9 3*-9784*9* 280*7--670 203020002* 8*376922*1 0*80900000 000655*02*
varia.   0211103221 0023042314 4350434142 01-4143534 25312--454 5131200001 0324443452 4242423242 1414454113

         5   :    15   :    25   :    35   :    44      : 52      : 62      : 72      : 82      :
HRAS     KLVVVGAGGV GKSALTIQLI QNHFVDEYDP TI-EDSYRKQ VVIDG--ETC LLDILDTAGQ EEYSAMRDQY MRTGEGFLCV FAINNTKSFE
DdRAS    KLVIVGGGGV GKSALTIQLI QNHFIDEYDP TI-EDSYRKQ VTIDE--ETC LLDILDTAGQ EEYSAMRDQY MRTGQGFLCV YSITSRSSFD
RASY1    KIVVVGGGGV GKSALTIQFI QSYFVDEYDP TI-EDSYRKQ VVIDD--KVS ILDILDTAGQ EEYSAMREQY MRTGEGFLLV YSVTSRNSFD
RASSP    KLVVVGDGGV GKSALTIQLI QSHFVDEYDP TI-EDSYRKK CEIDG--EGA VLDLLDTAGQ EEYSAMREQY MRTGQGFLLV YNITSRSSFD
TC21     RLVVVGGGGV GKSALTIQFI QSYFVTDYDP TI-EDSYTKQ CVIDD--RAA RLDILDTAGQ EEFGAMREQY MRTGEGFLLV FSVTDRGSFE
DRAS2    KLVVVGGGGV GKSAITIQFI QSYFVTDYDP TI-EDSYTKQ CNIDDIHNNL IF1VLDTAGQ EEFSAMREQY MRSGEGFLLV FALNDHSSFD
RRAS     KLVVVGGGGV GKSALTIQFI QSYFVSDYDP TI-EDSYTKI CSVDG--IPA RLDILDTAGQ EEFGAMREQY MRAGHGFLLV FAINDRQSFN
RAP1A    KLVVLGSGGV GKSALTVQFV QGIFVEKYDP TI-EDSYRKQ VEVDC--QQC MLEILDTAGT EQFTAMRDLY MKNGQGFALV YSITAQSTFN
DRAS3    KIVVLGSGGV GKSALTVQFV QCIFVEKYDP TI-EDSYRKQ VKVNE--RQC MLEIVNTAGT EQFTAMRNLY MKNGSD-SCW STRSRRNRRL
RAP2     KVVVLGSGGV GKSALTVQFV TGTFIEKYDP TI-EDFYRKE IEVDS--SPS VLEILDTAGT EQFASMRDLY IKNGQGFILV YSLVNQQSFQ
RSR1     KLVVLGAGGV GKSCLTVQFV QGVYLDTYDP TI-EDSYRKT IEIDN--KVF DLEILDTAGI AQFTAMRELY IKSGMGFLLV YSVTDRQSLE
RALA     KVIMVGSGGV GKSALTLQFM YDEFVEDYEP TK-ADSYRKK VVLDG--EEV QIDILDTAGQ EDYAAIRDNY FRSGEGFLCV FSITEMESFA

RHOY2    KLVIIGDGAC GKTSLLYVFT LGKFPEQYHP TV-FENYVTD CRVDG--IKV SLTLWDTAGQ EEYERLRPFS YSKADIILIG FAVDNFESLI
RAC1     KCVVVGDGAV GKTCLLISYT TNAFPGEYIP TV-FDNYSAN VMVDS--KPV NLGLWDTAGQ EDYDRLRPLS YPQTDVFLIC FSLVSPASYE
CDC42    KCVVVGDGAV GKTCLLISYT TNQFPADYVP TV-FDNYAVT VMIGD--EPY TLGLFDTAGQ EDYDRLRPLS YPSTDVFLVC FSVISPPSFE
TC10     KCVVVGDGAV GKTCLLMSYA NDAFPEEYVP TV-FDHYAVS VTVGG--KQY LLGLYDTAGQ EDYDRLRPLS YPMTDVFLIC FSVVNPASFQ
RHO1A    KLVIVGDGAC GKTCLLIVFS KDQFPEVYVP TV-FENYVAD IEVDG--KQV ELALWDTAGQ EDYDRLRPLS YPDTDVILMC FSIDSPDSLE
RHOY1    KLVIVGDGAC GKTCLLIVFS KGQFPEVYVP TV-FENYVAD VEVDG--RRV ELALWDTAGQ EDYDRLRPLS YPDSNVVLIC FSIDLPDSLE

RAB1A    KLLLIGDSGV GKSCLLLRFA DDTYTESYIS TIGVDFKIRT IELDG--KTI KLQIWDTAGQ ERFRTITSSY YRGAHGIIVV YDVTDQESFN
YPTM     KLLLIGDSSV GKSCFLLRFA DDSYVDSYIS TIGVDFKIRT VEVEG--KTV KLQIWDTAGQ ERFRTITSSY YRGAHGIIIV YDITDMESFN
SAS1     KLLLIGDSGV GKSCLLLRFS EDSFTPSFIT TIGIDFKIRT IELEG--KRI KLQIWDTAGQ ERFRTITTAY YRGAMGILLV YDVTDEKSFG
YPT2     KLLLIGDSGV GKSCLLLRFS EDSFTPSFIT TIGIDFKIRT IELDG--KRI KLQIWDTAGQ ERFRTITTAY YRGAMGILLL YDVTDKKSFD
SEC4     KILLIGDSGV GKSCLLVRFV EDKFNPSFIT TIGIDFKIKT VDING--KKV KLQLWDTAGQ ERFRTITTAY YRGAMGIILV YDVTDERTFT
RAB3A    KILIIGNSSV GKTSFLFRYA DDSFTPAFVS TVGIDFK-KT IYRND--KRI KLQIWDTAGQ ERYRTITTAY YRGAMGFILM YDITNEESFN
RAB2     KYIIIGDTGV GKSCLLLQFT DKRFQPVHDL TIGVEFGARM ITIDG--KQI KLQIWDTAGQ ESFRSITRSY YRGAAGALLV YDITRRDTFN
RAB4     KFLVIGNAGT GKSCLLHQFI EKKFKDDSNH TIGVEFGQKI INVGG--KYV KLQIWDTAGQ ERFRSVTTSY YRGAAGALLV YDITSRETYN
YPT3     KTVLIGDSGV GKSNLLMRFT RNEFNIESKS TIGVEFATRN IVLDN--KKI KAQIWDTAGQ ERYRAITSAY YRGAVGALIV YDITKQSSFD
Ara      KIVVIGDSAV GKSNLLSRYA RNEFSANSKA TIGVEFQTQS MEIEG--KEV KAQIWDTAGQ ERFRAVTSAY YRGAVGALVV YDITRRTTFE
RAB6     KLVFLGEQSV GKTSLITRFM YDSFDNTYQA TIGIDFLSKT MYLED--RTV RLQLWDTAGQ ERFRSLIPSY IRDSTVAVVV YDITNVNSFQ
RAB5     KLVLLGESAV GKSSLVLRFV KGQFHEFQES TIGAAFLTQT VCLDD--TTV KFEIWDTAGQ ERYHSLAPMY YRGAQAAIVV YDITNEESFA
RAB7     KVIILGDSGV GKTSLMNQYV NKKFSNQYKA TI-ADFLTKE VMVDD--RLV TMQIWDTAGQ ERFQSLGVAF YRGADCCVLV FDVTAPNTFK
TC4      KLVLVGDGGT GKTTFVKRHL TGEFEKKYVA TLGVEVHPLV FHTNR--GPI KFNVWDTAGQ EKFGGLRDGY YIQAQCAIIM FDVTSRVTYK
sites              PM1              G1        PM2                            PM3
```

```
            α3         L7        β5         L8                        α4        L9   β6     L10       α5
struc    HHHHHHHHHH HH              EEEEEE                                HHHHHHHHH  H      EEE              HH HHHHHHHHHH
water    509*228304 *16*4*---7 0100000063 67*7--------------68 0*4*908*40* 926-14324 000*97*208 *0053009628
varia.   3254444335 3-3543---4 3232110104 0044--------------25 14543244214 453-34020 3002445303 23054332425

         92       :102       109:      119:                      125      : 135      : 144      : 154      : 164
HRAS     DIHQYREQIK RVKDSD---D VPMVLVGNKC DLAA------ --------RT VESRQAQDLA RSYG-IPYIE TSAKTRQGVE DAFYTLVREIR
DdRAS    EIASFREQIL RVKDKD---R VPMIVVGNKC DLESD----- --------RQ VTTGEGQDLA KSFG-SPFLE TSAKIRVNVE EAFYSLVREIR
RASY1    ELLSYYQQIQ RVKDSD---Y IPVVVVGNKL DLENE----- --------RQ VSYEDGLRLA KQLN-APFLE TSAKQAINVD EAFYSLIRLVR
RASSP    EISTFYQQIL RVKDKD---T FPVVLVANKC DLEAE----- --------RV VSRREREQLA KSMH-CLYVE TSAKLRLNVE EAFYSLVRTIR
TC21     EIYKFQRQIL RVKDRD---E FPMILIGNKA DLDHQ----- --------RQ VTQEEGQQLA RQLK-VTYME ASAKIRMNVD QAFHELVRVIR
DRAS2    EIPKFQRQIL RVKDRD---E FPMLMVGNKC DLKHQ----- --------QQ VSLEEAQNTS RNLM-IPYIE CSAKLRVNVD QAFHELVRIVR
RRAS     EVGKLFTQIL RVKDRD---D FPVVLVGNKA DLESQ----- --------RQ VPRSEASAFG ASHH-VAYFE ASAKLRLNVD EAFEQLVRAVR
RAP1A    DLQDLREQIL RVKDTE---D VPMILVGNKC DLEDE----- --------RV VGKEQGQNLA RQWCNCAFLE SSAKSKINVN EIFYDLVRQIN
DRAS3    TICRTREQIL RVKDTD---D VPMVLVGNKC DLEEE----- --------RV VGKELGKNLA TQFN-CAFME TSAKAKVNVN DIFYDWSGRST
RAP2     DIKPMRDQII RVKRYE---K VPVILVGNKV DLESE----- --------RE VSSSEGRALA EEWG-CPFME TSAKSKTMVD ELFAEIVRQMN
RSR1     ELMELREQVL RIKDSD---R VPMVLIGNKA DLINE----- --------RV ISVEEGIEVS SKWGRVPFYE TSALLRSNVD EVFVDLVRQII
RALA     ATADFREQIL RVKEDE---N VPFLLVGNKS DLEDK----- --------RQ VSVEEAKNRA EQWN-VNYVE TSAKTRANVD KVFFDLMREIR

RHOY2    NARTKWADEA LRYCP----D APIVLVGLKK DLRQEAHFKE NATDE---MV PIEDAKQVAR AIGA-KKYME CSALTGEGVD DVFEVATRTSL
RAC1     NVRAKWFPEV RHHCP----S TPIILVGTKL DLRDDKDTIE KLKEKKLAPI TYPQGLALAK EIDS-VKYLE CSALTQRGLK TVFDEAIRAVL
CDC42    NVKEKWFPEV HHHCP----G VPCLVVGTQI DLRDDKVIIE KLQRQRLRPI TSEQGSRLAR ELKA-VKYVE CSALTQRGLK NVFDEAIVAAL
TC10     NVKEEWVPEL KEYAP----N VPFLLIGTQI DLRDDPKTLA RLNDMKEKPI CVEQGQKLAK EIGA-CCYVE CSALTQKGLK TVFDEAIIAIL
RHO1A    NIPEKWTPEV KHFCP----N VPIILVGNKK DLRNDEHTRR ELAKMKQEPV KPEEGRDMAN RIGA-FGYME CSAKTKDGVR EVFEMATRAAL
RHOY1    NVQEKWIAEV LHFCQ----G VPIILVGCKV DLRNDPQTIE QLRQEGQQPV TSQEGQSVAD QIGA-TGYYE CSAKTGYGVR EVFEAATRASL

RAB1A    NVKQWLQEID RYASE----N VNKLLVGNKC DLTTK----- --------KV VDYTTAKEFA DSLG-IPFLE TSAKNEKNVE QSFMTMAAEIK
YPTM     NVKQWLDEID RYAND----S VRNVLVGNKC DLAEN----- --------RA VDTSVAQAYA QEVG-IPFLE TSAKESINVE EAFLAMSAAIK
SAS1     NIRNWIRNIE QHATD----S VNKMLIGNKC DMAEK----- --------KV VDSSRGKSLA DEYG-IKFLE TSAKNSINVE EAFISLAKDIK
YPT2     NVRTWFSNVE QHASE----N VYKILIGNKC DCEDQ----- --------RQ VSFEQGQALA DELG-VKFLE ASAKTNVNVD EAFFTLAREIK
SEC4     NIKQWFKTVN EHAND----E AQLLLVGNKS DMET------ --------RV VTADQGEALA KELG-IPFIE SSAKNDDNVN EIFFTLAKLIQ
RAB3A    AVQDWSTQIK TYSWD----N AQVLLVGNKC DMEDE----- --------RV VSSERGRQLA DHLG-FEFFE ASAKDNINVK QTFERLVDVIC
RAB2     HLTTWLEDAR QHSNS----N MVIMLIGNKS DLESR----- --------RE VKKEEGEAFA REHG-LIFME TSAKTASNVE EAFINTAKEIY
RAB4     ALTNWLTDAR MLASQ----N IVIILCGNKK DLDAD----- --------RE VTFLEASRFA QENE-LMFLE TSALTGENVE EAFMQCARKIL
YPT3     NVGRWLKELR EHADS----N IVIMLVGNKT DLLHL----- --------RA VSTEEAQAFA AENN-LSFIE TSAMDASNVE EAFQTVLTEIF
Ara      SVGRWLDELK IHSDT----T VARMLVGNKC DLENI----- --------RA VSVEEGKALA EEEG-LFFVE TSALDSTNVK TAFEMVILDIY
RAB6     QTTKWIDDVR TERGS----D VIIMLVGNKT DLADK----- --------RQ VSIEEGERKA KELN-VMFIE TSAKAGYNVK QLFRRVAAALP
RAB5     RAKNWVKELQ RQASP----N IVIALSGNKA DLANK----- --------RA VDFQEAQSYA DDNS-LLFME TSAKTSMNVN EIFMAIAKKLP
RAB7     TLDSWRDEFL IQASPRDPEN FPFVVLGNKI DLENR----- --------QV ATKRAQAWCY SKNN-IPYFE TSAKEAINVE QAFQTIARNAL
TC4      NVPNWHRDLV RVCEN----- IPIVLCGNKV DLKDS----- --------KV KAKSIVFHRK KNL---QYYD ISAKSNYNFE KPFLWLARKLI
sites                                   G2                                         G3
```

B

```
HRAS                                MTEY...                                  QHKLRKLNPPDESGPGCMSCKCVLS      I
DdRAS                               MTEY...                                    KELKGDQSSGKAQKKKKQCLIL      IIa
RASY1                            MQGNKSTMTEY...                 110 a.a. RSKQSAEPQKNSSANARKEYSGGCCIIC      IIb
RASSP                               MRSTYLREY...                             RYNKSEEKGFQNKQAVCCVIC      IIb
TC21                             MAAAAGGRLRQEKY...                        KFQEQECPPSPEPTRKEKDKKGCHCVIF      I?
DRAS2                               MQMQTY...                              KFQIAQRPF-IEQDYKKKGKRKCCLM      IIb
RRAS     MSSGAASGTGRGRPRGGGPGPGDPPPSETH...                               KYQEQELPPSPPSAPRKKGGGCPCVLL      I?
RAP1A                               MREY...                                  RKTPVEKKKPKKKSCLLL      IIa
DRAS3                               MREY...                                  SRPRRNRRSRKVPCVLL      IIa
RAP2                                MREY...                                  YAAQPDKDDPCCSACNIQ      I
RSR1                                MRDY                           80a.a.SNRTGISATSQQKKKKKNASTCTIL      IIa
RALA                          MAANKPKGQNSLALH...                    ARKMEDSKEKNGKKKRKSLAKRIRERCCIL      IIb

RHOY2                            MSEKAVRR...                                 LMKKEPGANCCIIL      IIb
RAC1                                MQAI...                                  CPPPVKKRKRKCLLL      IIa
CDC42                               MQTL...                                  EPPVIKKSKKCTIL      IIa
TC10                          MPGAGRSSMAHGPGALML...                          TPKKHTVKKRIGSRCINCCLIT      IIb
RHO1A                               MAAIRK...                                QARRGKKKSGCLVL      IIa
RHOY1                            MSQQVGNSIRR...                           MGKSKTNGKAKKNTTEKKKKKCVLL      IIa

RAB1A                            MSSMNPEYDYLF...                      KRMGPGATAGGAEKSNVKIQSTPVKQSGGGCC      IIIa
YPTM                                MSNEFDYLF...                  KSKAGSQAALERKPSNVVQMKGRPIQQEQQKSSRCCST      IIb
SAS1                             MTSPATNKSAAYDYLI...                       KRMIDTPNEQPQVVQPGTNLGANNNKKKACC      IIIb
YPT2                                MSTKSYDYLI...                          KQKIDAENEFSNQANNVDLGNDRTVKRCC      IIIb
SEC4                          MSGLRTVSASSGNGKSYDSIM...                     EKIDSNKLVGVGNGKEGNISINSGSGNSSKSNCC      IIIb
RAB3A                      MASATDSRYGQKESSDQNFDYMF...                    EKMSESLDTADPAVTGAKQGPQLSDQQVPPHQDCAC      IV
RAB2                                MAYAYLF...EIYEKIQEGVFDINNEANGIKIGPQHAATNATHAGNQGGQQAGGGCC      IIIa
RAB4                                MSETYDFLF...          NKIESGELDPERMGSGIQYGDAALRQLRSPRRTQAPNAQECGC      IV
YPT3                                MCQEDEYDYLF...          RIVSNRSLEAGDDGVHPTAGQTLNIAPTMNDLNKKKSSSQCC      IIIb
Ara                              MSSDDEGREEYLF...           NNVSRKQLNSDTYKDELTVNRVSLVKDDNSASKQSSGFSCCSST      IIb
RAB6                                MSTGGDFGNPLRKF...                GMESTQDRSREDMIDIKLEKPQEQPVSEGGCSC      IV
RAB5                          MASRGATRPNGPNTGNKICQF...                    KNEPQNPGANSARGGGVDLTEPTQPTRNQCCSN      IIb
RAB7...                             MLL...                             KQETEVELYNEFPEPIKLDKNERAKASAESCSC      IV
TC4                                 MAAQGEPQVQF...                       GDPNLEFVAMPLSPHQKLSWTQLWQHSMSTT      -

Type I: xxxxxCxxxxxxCaaX;Type IIa: xxx (R,K) xxCaaX;Type IIb: similar, with extra Cys;
Type IIIa: xxxxxx(G)GGCC; Type IIIb: similar, but without Gly; Type IV: xxxxxxxxxxCAC.
(a=aliphatic)
```

FIGURE 3: Multiple sequence alignment of selected *ras*-like proteins. HRAS, DdRAS, etc. are abbreviated gene and protein names (see Figure 5 caption). A larger figure with all currently available sequences (full set as in Figure 5) is available from the authors on request. (A) Residues 5–164. β1, L1, α1, etc., are names of secondary structure elements as in Figure 1. Boldface columns: residues strictly conserved (variability 0). "struc": secondary structure according to DSSP (Kabsch & Sander, 1983), H = helix (α-helix or 3_{10}-helix), E = β strand. "water": solvent accessibility of a residue in a *ras* p21 monomer, calculated from the crystal structure, in units of the estimated number of water molecules in contact with this residue, equivalent to solvent-accessible surface area in units of 10 Å^2; residues with "water" = 0 are completely in the protein interior. "varia.": sequence variation at this sequence position on a scale of 0–5 (Sander & Schneider, 1991). Sequence position numbers refer to human H-*ras*; colons mark positions 10, 20, etc. and numbers 5, 15, 25, etc., are right-justified on the first residue in a block of ten. Hyphens indicate sequence gaps (insertions/deletions). Sites PM1, G1, etc., are as in Figure 2. For details of the roles of particular residues see Table I. (B) N-terminal sequences, preceding residue 5, and C-terminal sequences, following residue 164.

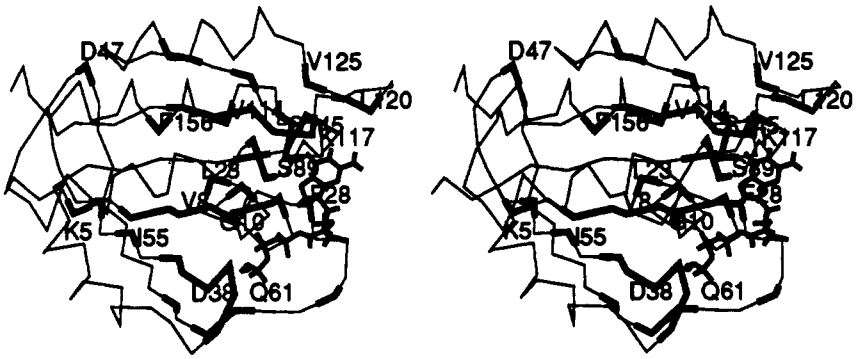

FIGURE 4: Stereoview of Cα trace of the high-resolution crystal structure of residues 1–166 of *ras* p21 protein determined by Pai et al. (1990). Very conserved residues are represented by thick lines; some of them are labeled. The GTP analogue is at the right-hand side. Note the cluster of most conserved residues surrounding the nucleotide-binding site and the fairly strong sequence conservation on the central β sheet and the helix α1/α5 contact site.

access to the catalytic site. In the adjacent position (Gly13), the Asp13 mutant is transforming (Fasano et al., 1984), but the Ser13 mutant is not. The three-dimensional structure and the biochemistry of Gly13 mutants have not yet been investigated. However, as Gly13 does have unusual backbone angles (ϕ = 78°, φ = 11°), the local structure of the phosphate binding loop is probably perturbed in most Gly13 mutants.

Given the oncogenic nature of mutations in position 12 and 13 in *ras*, the variation in the entire family of *ras*/*rho*/*ypt* proteins (Figure 3) in these positions is surprising: Gly-Gly in *ras* becomes Gly-Ala in *rho*, Ser-Gly, Thr-Gly, or no Gly in *rab* proteins. In spite of these deviations from *ras*, some *rab* proteins such as *rab3* (Ser-Ser) have a GTPase activity comparable to or even higher than p21 *ras* (Zahraoui et al.,

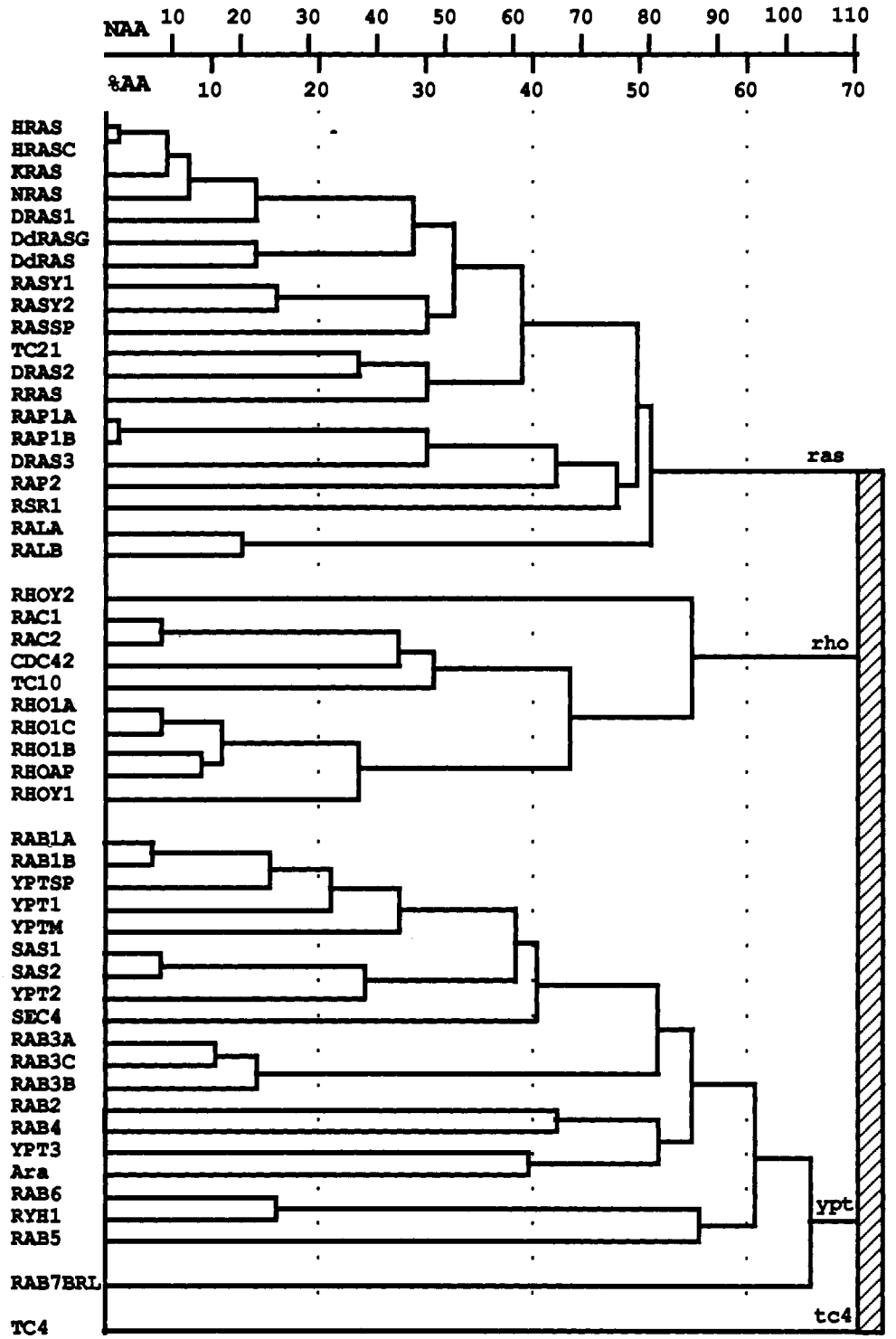

FIGURE 5: Evolutionary tree of *ras*-like protein sequences. Similarity between any two proteins can be read off the graph approximately by looking up NAA or AA at the leftmost tree node the proteins share. NAA: number of nonidentical amino acids in the region 5–164. %AA: percent nonidentical amino acids. For example, H-*ras* and Dd-*ras*-G are about 29% different in sequence, while H-*ras* and *tc4* are about 70% different. The bar on the right reflects uncertainty of tree topology at the root level for the four subfamilies *ras*/*rho*/*ypt*/*tc4*. The vertical distance between two proteins has no precise meaning. Sequence references follow. *ras* subfamily: H-*ras* (Capon et al., 1983), K-*ras* (McGrath et al., 1983), N-*ras* (Taparowsky et al., 1983), H-*rasC* (Westaway et al., 1986), D-*ras1* (Neuman-Silberberg, 1984), *rasY1* (DeFeo-Jones et al., 1983), *rasY2* (Powers et al., 1984), *rasSP*, (Fukui et al., 1985), Dd-*rasG* (Robbins et al., 1989), Dd-*ras* (Reymond et al., 1984), *tc21* (Drivas et al., 1990), R-*ras* (Lowe et al., 1987a,b), D-*ras2* (Mozer et al., 1985), *rap1A* (Pizon et al., 1988b; Kawata et al., 1988), K-*rev1* (Kitayama et al., 1989; Nagata et al., 1989), *rap1B* (Pizon et al., 1988a; Lapetina et al., 1989), Apl-*ras* (Swanson et al., 1986), *rap2* (Pizon et al., 1988b), *rsr1* (Bender & Pringle, 1989), D-*ras3* (Schejter & Shilo, 1985), *ralA* (Chardin & Tavitian, 1989; Bhullar et al., 1990), and *ralB* (Chardin & Tavitian, 1989). *rho* subfamily: *rho1A* (Yeramian et al., 1987), *rho1B* (Chardin et al., 1988), *rho1C* (Chardin et al., 1988), *rho*-Apl (Madaule & Axel 1985), *rhoY1* (Madaule et al., 1987), *rac1* (Didsbury et al., 1989), *rac2* (Didsbury et al., 1989), G25K (Polakis et al., 1989a; Evans et al., 1986), *cdc42* (Johnson & Pringle 1990), *tc10* (Drivas et al., 1990), and *rhoY2* (Madaule et al., 1987). *ypt*/*rab* subfamily: *rab1A* (Touchot et al., 1987), *rab1B* (Vielh et al., 1989), *yptm* (Palme et al., 1988), *ypt1* (Gallwitz et al., 1983), *yptSP* (Fawell et al., 1989), *rab2* (Touchot et al., 1987), *rab3A* (Touchot et al., 1987; Matsui et al., 1988), *rab3B* (Zahraoui et al., 1988), *rab3C* (Matsui et al., 1988), *rab4* (Touchot et al., 1987), *rab6* (Zahraoui et al., 1989), *ryh1* (Hengst et al., 1990), *rab5* (Zahraoui et al., 1989), *ypt3* (Miyake & Yamamoto, 1990), *ara* (Matsui et al., 1989), *sas1* (Saxe & Kimmel, 1988), *sas2* (Saxe & Kimmel, 1988), *ypt2* (Haubruck et al., 1990), *sec4* (Salminen & Novick, 1987), and *rab7* (Bucci et al., 1988). *tc4* possible subfamily: *tc4* (Drivas et al., 1990).

```
                PM1                    PM2                PM3                G2
SPR VVTFCGVNGVGKSTNLAKISF          <75>          VVLVDTAGRMQ <55> IDGIVLTKFDT
SP5 VIMFVGLQGSGKTTTCSKLAY          <62>          IIIVDTSGRHK <45> VASVIFTKLDG
Era FIAIVGRPNVGKSTLLNKLLG          <27>          AIYVDTPGLHM <49> PVILAVNKVDN
Lep NFSIIAHIDHGKSTLSDRIIQ          <46>          LNFIDTPGHVD <41> EVVPVLNKIDL
Gαt KLLLLGAGESGKSTIVKQMKI         <140>          FRMFDVGGQRS <56> SIVLFLNKKDV
EFT NVGTIGHVDHGKTTLTAAITT  <26> ITINTSH  <9> YAHVDCPGHAD <42> YIIVFLNKCDM
Sar KLLFLGLDNAGKTTLLHMLKN   <7> PTWHPTS  <9> FTTFDLGGHIQ <46> FFVILGNKIDA
ARF RILMVGLDAAGKTTILYKLKL   <7> PTIGFNV  <9> FTVWDVGGQDK <46> VLLVFANKQDL
Ras KLVVVGAGGVGKSALTIQLIQ   <8> PTIEDSY <12> LDILDTAGQEE <46> PMVLVGNKCDL
          •    •                •                  •                  •
Num      10   16                35                 57                119
Ptn  eeeeeGxxxxGKs              xTx               eeeeDxxG           eeeeenKxD
Str EEEEE_____HHHHHHHHHH        ___EEEE           EEEEE_____         EEEEEE_____
      β1        α1               β2                 β3                 β5
```

FIGURE 6: Conserved sequence regions common to a wider family of G-proteins, distantly related to the *ras/rho/ypt* family. Only a few representative members of these sequence families are shown. The PM2 box at T35 as well as the G1 and G3 boxes may have analogues in all sequences, but they are not shown except for PM2, where it could be assigned unambiguously. Ptn, characteristic sequence pattern in terms of conserved residues or residue properties; x = any residue type; s = S or T; n = N, C, T, L, or I; eeee = run of residues with clear average β-strand preference; str, known secondary structure of *ras* and EF-Tu; H = α-helix; E = β-strand; num, residue numbers in human *ras*. <75>, etc., are sequence gaps. EMBL/Swissprot sequence identifiers are as follows: SPR = SRPR$HUMAN, SP5 = SRP5$MOUSE, ERA = ERA$ECOLI, LEP = LEPA$ECOLI, GAT = GBT1$HUMAN, EFT = EFTU$ECOLI, ARF = ARF3$HUMAN, and RAS = RASH$HUMAN. In SP5, there are two possible assignments of the DxxG and nKxD motifs—we have chosen the one more consistent with β-strand preferences and sequence gaps.

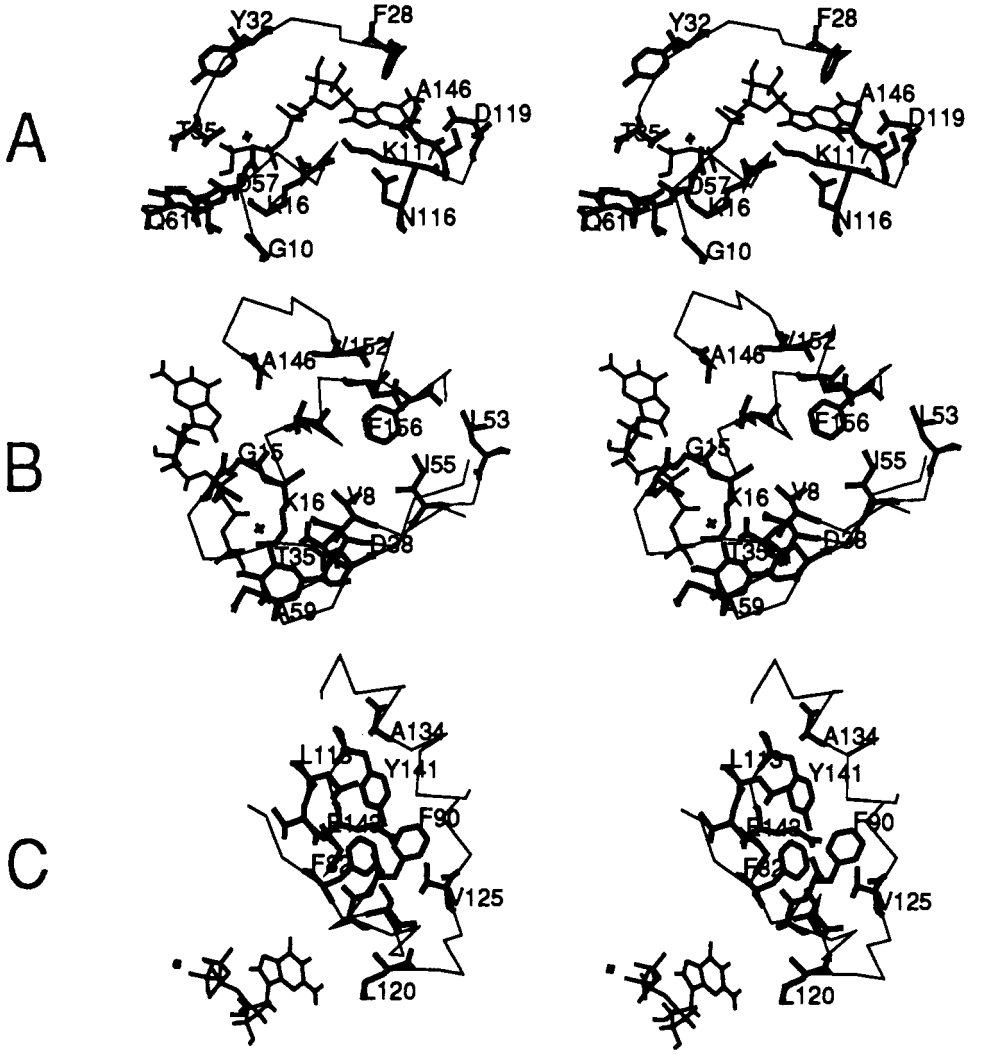

FIGURE 7: Stereoviews of conserved residues in the *ras* p21 crystal structure. The bound nucleotide and selected side chains are in all-atom detail; for other selected residues, C(α) atoms are connected by a virtual bond for simplicity. (A) Conserved residues in contact with the bound nucleotide. (B) Cluster of mutually contacting conserved residues (C1 in Table I). (C) Cluster of mutually contacting conserved residues (C2 in Table I).

1989); the same is true for *rho* proteins with Ala13 (Garret et al., 1989). This suggests that the local structure of the nucleotide-binding loop around the β,γ-phosphates is slightly different in *ypt/rab* and *rho* proteins. In addition, the mechanism of GTP hydrolysis may be different in detail from that proposed for *ras* p21 (Pai et al., 1990).

Another conserved and apparently crucial residue in the PM site is Thr35 (PM2), which in the high-resolution crystal structure is seen to participate in the octahedral coordination of the Mg^{2+} ion and in the stabilization of the γ-phosphate. Thr35 may also be involved in activation of H_2O for GTP hydrolysis (Pai et al., 1990). It also has a role in the conformational transition from the GTP-bound form to the GDP-bound form of ras p21 (Milburn et al., 1990; Schlichting et al., 1990), since removal of the γ-phosphate and loss of hydrogen bonds to Thr35 allows a shift of loop L2, which is involved in the interaction with the effector.

The second highly conserved region is the DTAGQE motif around position 60 (PM3 site), with Gln61 replaced by Thr in the *rap*/D-*ras3* proteins and with Asp57 and Gly60 apparently essential for all G-proteins. The single exception is D-*ras3*, which has Asn57. Formally, the variation in the *ras* family is [D,N]TAG[Q,T,I][E,A] (residues in square brackets are alternatives at one position). Oligonucleotides corresponding to DTAGQE have been used to detect a number of *ras* sequences in cDNA libraries (Touchot et al., 1987; Chavrier et al., 1990). Asp57 is coordinated to Mg^{2+} via a water molecule in the triphosphate structure (Pai et al., 1990). After GTP hydrolysis it serves as the second negatively charged lignad of Mg^{2+} in the *ras*–GDP complex and is thus an important element of the conformational change as seen in time-resolved crystallographic studies (Schlichting et al., 1990). Likewise, Gly60 is hydrogen-bonded to the γ-phosphate and also involved in triggering a conformational change when the γ-phosphate is hydrolyzed.

The DTAGQE motif overlaps with the most flexible region of *ras* p21 (residues 61–65). In the high-resolution structure, alternate conformations of these residues help to explain the electron density. It has been postulated that Gln61 is involved in catalysis because the side chain of Gln61 is able to adopt a conformation in which it is close to a water molecule thought to attack the γ-phosphate (Pai et al., 1990). Substitution of Gln61 by another residue indeed reduces GTPase activity and renders the protein oncogenic (Der et al., 1986a). The *rap1A* and *rap2* proteins have Thr61 instead of Gln61, suggesting that these proteins have a different mechanism for the GTPase cleavage step. Subunits of heterotrimeric G-proteins also have Gln in this position, like *ras*, and it has been found that the substitution of this Gln by Arg significantly reduces the GTPase rate (Graziano & Gilman, 1989; Masters et al., 1986). In elongation factor Tu the corresponding residue is His84, also involved in GTP hydrolysis but not in substrate binding (Cool & Parmeggiani, 1991).

Turning from phosphate/Mg binding to binding of the guanine base, residue Phe28 (site G1, loop L2) is highly conserved, has its ring approximately perpendicular to the guanine base in the GDP- and the GTP-bound form, and is well packed against the aliphatic stem of the conserved Lys147 (Figure 7A). Phe28 is replaced by Tyr in several *rab* genes, requiring at least some adjustment of the interaction between residue 28 and the guanine base to make room for the extra hydroxyl group.

The NKxDL motif, residues 116–120 (G2 site), is located at the end of strand β5 and the beginning of loop L8. The actual variation observed so far in the *ras*/*rho*/*ypt* family is [N,T,I,C,L][K,Q]xD[L,M,C,I]. The Asn116 side chain in *ras* p21 interacts both with the hydroxyl group of the Thr144 side chain and with the main-chain HN of Val14 in the phosphate-binding loop, forming a bridge between three of the nucleotide-binding loops; Asn116 also interacts weakly with the N7 atom of the guanine ring. Thr116 in place of Asn116 can only partially fulfill such a multiple role, suggesting differences in nucleotide-binding properties and in local architecture when Asn116 is absent, as in most *rho* proteins. The main role of Lys117 appears to be hydrophobic interaction between its aliphatic side chain and the aromatic ring of the guanine base. Lys117 interacts only weakly with O1′ of the ribose. It seems plausible that the side chain of Gln can perform the same functions of hydrophobic and polar interaction in this position, as in *cdc42* and *tc10*. Asp119 is the only totally conserved residue in this motif. Its side chain forms a bifurcated hydrogen bond to the endocyclic NH and the exocyclic NH_2 group of the base. It also interacts with the side chain of the strictly conserved Ser145. Numerous mutations of residues in the NKxD motif have been engineered into *ras* p21 and other G-proteins. The resulting mutants have greatly increased dissociation rate constants and, where it has been investigated, produced a different specificity for nucleotides (Sigal et al., 1986; Feig et al., 1986; Der et al., 1986b). The NKxD motif is conserved in all G-domains (Figure 6).

The ExSA motif or G3 site, starting at E143, is also conserved in all proteins of the ras family. The observed variation is [Y,F,H][I,V,L,H,F,M]E[T,S,A,C]SA[K,L,M], starting at Y141. The three-dimensional structure of p21 shows no direct involvement of these side chains in binding of the guanine base, but there is indirect interaction: the main-chain NH of A146 makes a strong hydrogen bond with O6 of the base, and the side chain of Ser145 forms a hydrogen bond to the side chain of Asp119. Ala146 seems to be conserved to steric reasons (see below). The aliphatic side chain of K147 makes a strong aliphatic interaction with F28 that in turn interacts with the base. Replacement of this Lys by Leu or Met apparently preserves this hydrophobic interaction. Glu143 has several other contacts with residues in the NKxD and ExSA loops, but it isnot clear why it should be strongly conserved. The ExSA motif thus has a helper function in the binding or dissociation of the guanine base. It is not obviously conserved in the wider class of all G-proteins.

Conserved Residues Not Involved in Nucleotide Binding

One can assume that other conserved residues either are involved in maintaining the three-dimensional structure or participate in a common functional property other than nucleotide binding. Generally speaking, conserved residues involved in many contacts with other residues probably are important in maintaining the structure, e.g., residues at the helix–sheet interface. In contrast, very solvent-accessible conserved residues are likely to be involved in an important interaction with an external molecule (Valencia et al., 1991), e.g., with GAP protein. Looking at the distribution of conserved residues (where conserved is taken as variability 0 or 1 in Figure 3), there is a remarkable concentration near the nucleotide-binding pocket and in the adjacent helix–sheet interfaces (Figure 4). These residues can be grouped into two clusters, one in the interface between helices α1 and α5 and strands β1 and β4 of the β sheet, called cluster C1 (Figure 7B), and one between helices α3 and α4 and strands β4, β5, and β6 on the other side of the β sheet, called cluster C2 (Figure 7C). Almost all conserved residues in the protein interior, i.e., residues rich in intraprotein contacts, are involved in contacts with other conserved residues, e.g., F156 is in contact with I55.

Conservation of residues in these clusters probably reflects structural requirements, and they may be folding nuclei. A few of the conserved residues in the nucleotide-binding pocket

Table I: Role of Conserved Residues in *ras* p21[a]

Cluster C1: β1 (V8, V9), L1 (V14, *G15*), α1 (K16, L19, L23), L2 (*T35*), β2 (D38), β3 (L53, I55, D57, *T58*), L4 (A59, *G60*), L10 (*A146*), α5 (V152, F156)	
V8, V9	fix position of K16, which contacts GNP
V14	contact with cluster 2
G15	part of the Mg/phosphate site PM1
K16	many contacts, β1 (V8, V9), β3 (D56, T58), L1 (V14), L4 (G60), part of site PM1
L19, L23	contact with α5 (V152, F156, L19) and with cluster 2 (N116)
T35	in site PM2
D38	only conserved residue in β2, very exposed
I55	in contact with α5 (F156)
D57, *T58*, *A59*, and *G60*	part of Mg/phosphate site PM3, T58 contacts V8
A146	part of site G3
V152, F156	most conserved resiudes on α5, strong contact with α1 and β3
Cluster C2: β4 (F82), L6 (I84), α3 (S89, F90), β5 (L113, V114, G115), L8 (L120, V125), α4 (A134), β6 (Y141, E143)	
F82, I84	part of a hydrophobic pocket around the G base, F82 contacts α3, β5, β6 and L8
S89	in contact with L1 (V14) of cluster 1
F90	α/β contact with β4 (F82)
L113, V114	hydrophobic pocket at G base
G115	side chain here would clash with conserved F82
L120	contact with K117 in G2 and with β4
V125	contacts β4 and α3
A134	strong α/β contact with β5 (L113) and β6 (Y141)
Y141	contacts β4, β5, and α4
E143	H-bond to Y141, salt bridge to R123 (conserved in *ras* subfamily only)
Not in Contact with the Clusters: β1 (K5, V7, *G10*), L2 (*F28*, Y32, I36), L3 (D47), L4 (*Q61*, E62), L8 (*N116*, *K117*, *D119*), L10 (*S145*)	
G10	part of PM1
F28	contact with guanine base (site G1)
I36, E62	mutual contact
Q61	part of PM3
N116, *K117*, and *D119*	contacts guanine base (site G2), N119 also contacts S145
S145	contacts D119 in G2; part of site G3
K5	head group exposed, may make functional external contact
V7	on β1, contact with Y71 on α2
Y32	possibly interaction with GAP
D47	β hairpin β2/β3, very exposed, possible functional contact

[a] All conserved residues are listed (variability 0 or 1, values from Figure 3). Residues involved in nucleotide binding (sites PM and G, Figure 6) are shown in italic type. α, β, and L (helix, β strand, and loop) refer to secondary structure elements (Figure 1).

(Figure 7A) are part of cluster C1, notably G15, K16 of site PM1, and T35 of site PM2. The general impression is that the two clusters form a stable structural core spatially adjacent to the substrate-binding site and that this core precisely determines the location of the liganding residues.

The properties and putative roles of all residues with low variability are summarized in Table I. Residues for which the reasons for conservation are not already obvious require further comment. The first of these is K5, on the edge of cluster C1 and partly exposed to solvent, which has no obvious structural role. Another is E62 in loop L4. Conceivably Lys5 or Glu62 could be involved in interaction with a functionally important external partner specific to the *ras/rho/ypt* family, as they are not conserved in the larger class of G-domains.

Other cases of conservation difficult to understand include A59, of the DTAG motif, which makes no obvious interaction with other residues or GTP. It is in contact with the side chains of Y64 and E62 and also not far from a water molecule that is believed to be involved in GTP hydrolysis. Introduction of larger side chains for A59 may lead to steric interference. In fact, the mutation A59T has a dramatic influence on nucleotide dissociation and GTP hydrolysis rates (John et al., 1988). With a minor conformational change the hydroxyl side chain of T59 could come close enough to the γ-phosphate to act as a nucleophile in an autophosphorylation reaction (John et al., 1988). In the same motif, T58 makes no GTP contact yet is very conserved. Its side chain, however, makes a hydrogen bond to the main-chain O of V8; this may strengthen the connection between the β strands just before the two phosphate-binding loops L1 and L4. Finally, strictly conserved A146 in cluster C1, of the ExSAK motif, is near site G3 and its main-chain carbonyl interacts with O6 of the base. Perhaps substitution by a larger residues would lead to steric hindrance, disrupting this hydrogen bond.

In summary, most of the conserved residues that are not involved in nucleotide binding appear to be part of the structural core of *ras* proteins. For some residues the rationale for conservation is unclear. Such residues are interesting targets for mutation experiments. Some further insight regarding these residues can be gained by comparing the conservation of structurally equivalent residues in a related family of G-domains, that of the elongation factor Tu (Valencia et al., 1991).

The N-Terminus

There is considerable variation of length and sequence in the N-terminal region, which we take to end at the conserved Lys5, near the beginning of β strand 1 (Figure 3A). While these sequences are usually very similar in subbranches of the evolutionary tree (Figure 5), they can differ widely from one subbranch to the other. In H-*ras* p21, the N-terminal extension is very short and takes full part in the β sheet. In other proteins, such as in R-*ras*, *rab3*, *rab5*, or *sec4*, there are up to 30 residues, enough for an additional subdomain. These N-terminal extensions would be located in the general vicinity of the C-terminus and may be involved in direct physical interaction with the C-terminal extension, sharing in its functional role.

The C-Terminus and Cysteine Motifs

The region that can be unambiguously aligned ends at R164, at the end of the C-terminal helix in the crystal structure of truncated H-*ras*. The sequences after this position vary greatly in length and sequence but have approximately the same length in each of the three main subbranches. The shortest C-terminal extensions are found in the *rho* branch (14–17 residues), so in these proteins the major globular domain is expected to be closer to the membrane. The proteins of the *ras* branch have extensions of 18–30 amino acids, while in the *ypt/rab* branch their length ranges from 27 to 47 amino acids. An extension of about 50 residues can represent an additional small domain of the protein. In yeast *ras1* and *ras2*, however, the C-terminal domains are nearly as large as the guanine nucleotide binding domain.

The sequences can be aligned on conserved Cys residue motifs at the C-terminus (Figure 3B), known to be involved in covalent modification and/or membrane attachment. Particular C-terminal sequences may determine association with specific membranes, possibly through interaction with membrane-bound proteins. In the *rab/ypt/sec4* subfamily, the C-terminus may determine which particular intracellular

transport pathway the protein is involved in (Chavrier et al., 1990).

The conserved cysteine motif at the C-terminus (Figure 3B) is Caax, CCax, GGCC, or xCxC, where x is any amino acid and a is an aliphatic residue. In an apparently complicated series of reactions, the cysteine in the Caax sequence of *ras* proteins is farnesylated, the three following amino acids are protetolytically removed, and the now C-terminal farnesylated cysteine is carboxymethylated. Farnesylation occurs rapidly after protein synthesis and is irreversible (Gutierrez et al., 1989; Hancock et al., 1989; Schafer et al., 1989). The order of the posttranslational modifications is not precisely known. Proteolysis is not the first step since a protein ending at Cys186 cannot undergo farnesylation. *rab* proteins ending with a Cys residue are probably not farnesylated. However, they can probably be carboxymethylated and undergo additional fatty acylation in the various vesicles or organelle membranes. In *ras* proteins, additional cysteins close to the C-terminal end can become palmitylated. It is likely that in other *ras*-like proteins an accessible Cys may become fatty acylated if it is located close enough to the membrane.

Several proteins, mainly in the *rab* branch, have a C-terminal cysteine motif that is significantly different from the Caax motif (Figure 3B). It is not precisely known what changes can be tolerated relative to the Caax motif without affecting farnesylation, carboxymethylation, and clipping. What kinds of modifications occur in proteins ending with CCIL like *ral*, CCIIC like yeast *ras1*, or CaC like several *rab* proteins? A variant has already been found for *rab3* (Fischer-von-Mollard et al., 1990) that appears to be modified by the addition of a hydrophobic group sensitive to hydroxylamine treatment and, in a second step, to acquire a higher avidity for membranes and to become resistant to hydroxylamine. This is in contrast to *ras*, where the first step, farnesylation, is not sensitive to hydroxylamine while the second step, palmitylation, is sensitive.

The precise study of the posttranslational modifications on these proteins should provide interesting insights into the requirements and specificity of the modifying enzymes involved and might lead to the discovery of different mechanisms. Interestingly, the α subunits of heterotrimeric G-proteins ending with a [D,E]CGLx sequence are not farnesylated. In farnesylated proteins E or D before the C and G after the C have not been found, suggesting that their presence might somehow impair recognition by the farnesyl transferase. Only *tc4* protein has no cysteine in the C-terminal end, underscoring its classification as the first member of a new subfamily of *ras*-like proteins.

Relation to Other G-Protein Families

Further insight can be gained from comparing *ras* proteins to the largest possible known family of G-proteins. On the basis of the crystal structure of one other G-domain, that of elongation factor Tu (LaCour et al., 1985; Jurnak, 1985), and exhaustive sequence comparisons, it appears that G-domains can be characterized by a minimal sequence pattern, as shown in Figure 6. This pattern was derived from an alignment of *ras* with other known G-proteins, representing different families: ADP-ribosylation factor, ER-to-Golgi transport protein Sar1p, bacterial gene product Era, bacterial gene product LepA (cotranscribed with signal peptidase), signal recognition particle receptor subunit, signal recognition particle 54-kDa subunit, a G-protein α subunit, and elongation factor Tu.

The nucleotide-binding GxxxGKs, DxxG, and nKxD motifs [as in Dever et al. (1987)] are strongly conserved in the larger family. The β strands immediately adjacent to these motifs also appear to be present in each of these proteins, on the basis of the observed structural preference of runs of 3–5 residues, like ILMV, FTVW, or MVLVG (protein ARF). Residue T35 of H-*ras*, essential for Mg and γ-phosphate coordination in *ras* proteins, appears in approximately the expected position in foru of the proteins (Figure 6). Identification of this single residue in the other five proteins is ambiguous, especially when the sequence gaps are much larger than in H-*ras*. The ETSAK motif, conserved in the *ras* family, is not conserved in the more general family.

What do the conserved sequence patterns tell us? Which properties are conserved in all G-domains? There is insufficient biochemical data to describe in detail the similarities and differences. However, what appears to be conserved is the overall three-dimensional structure of each domain, i.e., the structural core with the central β sheet sandwiched between α helices, and the location of the nucleotide-binding loops at the C-terminal end of strands β1, β3, and β5, as well as the order in which the secondary structure elements are threaded.

It appears likely that most of the proteins in this wider class are GTPases with the common biological function of a timed switch. The switch is turned on (or off) as a result of encounter with specific macromolecular patterns, often other proteins, and the result of switching is modified inteaction with other molecules. It will be very interesting to unravel the details of how this protein switch is used in different functional cellular contexts, ranging from growth control to interorganelle traffic.

Added in Proof

Two reviews about the more general class of GTP-binding proteins have appeared recently (Bourne et al., 1990a,b).

Acknowledgments

We thank Gert Vriend for the use of the WHAT IF program, Reinhard Schneider for the database HSSP, Martin Vingron for calculation of an earlier version of the phylogenetic tree, Ilme Schlichting for discussion of *ras* structures, and Ken Holmes for initiating the *ras* project. We apologize to authors of papers containing sequences whose work we were not able to cite for space reasons. A.V. was supported by an EMBO fellowship.

References

Ahnn, J., March, P. E., Takiff, H. E., & Inouye, M. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 8849–8853.

Barbacid, M. (1987) *Annu. Rev. Biochem. 56*, 779–827.

Bender, A., & Pringle, J. R. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 9976–9980.

Bhullar, R. P., Chardin, P., & Haslam, R. J. (1990) *FEBS Lett. 260*, 48–52.

Bos, J. L. (1988) *Mutat. Res. 195*, 255.

Bourne, H. R., Sanders, D. A., & McCormick, F. (1990a) *Nature 348*, 125–132.

Bourne, H. R., Sanders, D. A., & McCormick, F. (1990b) *Nature 349*, 117–127.

Bucci, C., Frunzio, R., Chiariotti, L., Brown, A., Rechler, M., & Bruni, C. (1988) *Nucleic Acids Res. 16*, 9979–9993.

Capon, D. J., Che, E. Y., Levinson, A. D., Seeburg, P. H., & Goeddel, D. V. (1983) *Nature 302*, 33–37.

Chardin, P. (1988) *Biochimie 70*, 865–868.

Chardin, P., & Tavitian, A. (1989) *Nucleic Acids Res. 17*, 4380.

Chardin, P., Madaule, P., & Tavitian, A. (1988) *Nucleic Acids Res. 16*, 2717.

Chavrier, P., Vingron, M., Sander, C., Simons, K., & Zerial, M. (1990) *Mol. Cell. Biol. 10*, 6578–6585.

Cool, R. H., & Parmeggiani, A. (1991) *Biochemistry 30*, 362–366.
Defeo-Jones, D., Scolnick, E. M., Koller, R., & Dhar, R. (1983) *Nature 306*, 707–709.
Dever, T. E., Glynias, M. J., & Merrick, W. C. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 1814–1818.
deVos, A., Tong, L., Milburn, M., Matias, P., Jancarik, J., Noguchi, S., Nishimura, S., Miura, K., Ohtsuka, E., & Kim, S.-H. (1988) *Science 239*, 888–893.
Der, C. J., Finkel, T., & Cooper, G. M. (1986a) *Cell 44*, 167–176.
Der, C. J., Pan, B. T., & Cooper, G. M. (1986b) *Mol. Cell. Biol. 6*, 3291–3294.
Didsbury, J., Weber, R. F., Bokoch, G. M., Evans, T., & Snyderman, R. (1989) *J. Biol. Chem. 264*, 16378–16382.
Dreusicke, D., & Schulz, G. E. (1986) *FEBS Lett. 208*, 301–304.
Drivas, G. T., Shih, A., Coutavas, E., Rush, M. G., & D'-Eustachio, P. (1990) *Mol. Cell. Biol. 10*, 1793–1798.
Evans, T., Brown, M., Fraser, E., & Northup, J. (1986) *J. Biol. Chem. 261*, 7052–7059.
Fasano, O., Aldrich, T., Tamanoi, F., Taparowsky, E., Furth, M., & Wigler, M. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 4008–4012.
Fawell, E., Hook, S., & Armstrong, J. (1989) *Nucleic Acids Res. 17*, 4373.
Feig, L., Pan, B.-T., Roberts, T., & Cooper, C. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 4607–4611.
Felsenstein, J. (1981) *J. Mol. Evol. 17*, 368–376.
Fischer-von-Mollard, G., Mignery, G. A., Baumert, M., Perin, M. S., Hanson, T. J., Burger, P. M., Jahn, R., & Sudhof, T. C. (1990) *Proc. Natl. Acad. Sci. U.S.A. 87*, 1988–1992.
Fitch, W. M., & Margoliash, E. (1967) *Science 155*, 279–284.
Fukui, Y., & Kaziro, Y. (1985) *EMBO J. 4*, 687–691.
Gallwitz, D., Donath, C., & Sander, C. (1983) *Nature 306*, 704–707.
Garrett, M., Self, A., vanOers, C., & Hall, A. (1989) *J. Biol. Chem. 264*, 10–13.
Gay, N. J., & Walker, J. E. (1983) *Nature 301*, 262–264.
Graziano, M., & Gilman, A. G. (1989) *J. Biol. Chem. 264*, 15467–15474.
Gutierrez, L., Magee, A., Marshall, C., & Hancock, J. (1989) *EMBO J. 8*, 1093–1098.
Hall, A. (1990) *Nature 249*, 635–640.
Hancock, J. F., Magee, A. I., Childs, J. E., & Marshall, C. J. (1989) *Cell 57*, 1167–1177.
Haubruck, H., Engelke, U., Mertins, P., & Gallwitz, D. (1990) *EMBO J. 9*, 1957–1962.
Hengst, L., Lehmeier, T., & Gallwitz, D. (1990) *EMBO J. 9*, 1949–1955.
John, J., Frech, M., & Wittinghofer, A. (1988) *J. Biol. Chem. 263*, 11792–11799.
Johnson, D. I., & Pringle, J. R. (1990) *J. Cell Biol. 111*, 143–152.
Jurnak, F. (1985) *Science 130*, 32–36.
Kabsch, W., & Sander, C. (1983) *Biopolymers 22*, 2577–2637.
Kawata, M., Matsui, Y., Kondo, J., Hishida, T., Teranishi, Y., & Takai, Y. (1988) *J. Biol. Chem. 263*, 18965–18971.
Kikuchi, A., Sasaki, T., Araki, S., Hata, Y., & Takai, Y. (1989) *J. Biol. Chem. 264*, 9133–9136.
Kitayama, H., Sugimoto, Y., Matsuzaki, T., Ikawa, Y., & Noda, M. (1989) *Cell 56*, 77–84.
Krengel, U., Schlichting, I., Scherer, A., Schumann, R., Frech, M., John, J., Kabsch, W., Pai, E. F., & Wittinghofer, A. (1990) *Cell 62*, 539–548.
LaCour, T. F. M., Nyborg, J., Thirup, S., & Clark, B. F. C. (1985) *EMBO J. 4*, 2385–2388.
Lapetina, E. G., Lacal, J. C., Reep, B. R. ,& Vedia, L. M. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 3131–3135.
Lowe, D., & Goeddel, D. (1987a) *Mol. Cell. Biol. 7*, 2845–2856.
Lowe, D., Capon, D., Delwart, E., Sakaguchi, A., Naylor, S., & Goeddel, D. (1987b) *Cell 48*, 137–146.
Madaule, P., & Axel, R. (1985) *Cell 41*, 31–40.
Madaule, P., Axel, R., & Myers, A. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 779–783.
March, P. E., & Inouye, M. (1985) *J. Biol. Chem. 260*, 7206–7213.
Masters, S. B., Stroud, R. M., & Bourne, H. R. (1986) *Protein Eng. 1*, 47–54.
Matsui, M., Sasamoto, S., Kunieda, T., Nomura, N., & Ishizaki, R. (1989) *Gene 76*, 313–319.
Matsui, Y., Kikuchi, A., Kondo, J., Hishida, T., Teranishi, Y., & Takai, Y. (1988) *J. Biol. Chem. 263*, 11071–11074.
McGrath, J. P., Capon, D. J., Smith, D. H., Chen, E. Y., Seeburg, P. H., Goeddel, D. V., & Levinson, A. D. (1983) *Nature 304*, 501–506.
Milburn, M. V., Tong, L., deVos, A. M., Brunger, A., Yamaizumi, Z., Nishimura, S., & Kim, S.-H. (1990) *Science 247*, 939–945.
Miyake, S., & Yamamoto, M. (1990) *EMBO J. 9*, 1417–1422.
Mozer, B., Marlor, R., Parkhurst, S., & Corces, V. (1985) *Mol. Cell. Biol. 5*, 885–889.
Nagata, K., Itoh, H., Katada, T., Takenaka, K., Ui, M., Kaziro, Y., & Nozawa, Y. (1989) *J. Biol. Chem. 264*, 17000–17005.
Neuman-Silberberg, F. S., Schejter, E., Hoffmann, F. M., & Shilo, B. Z. (1984) *Cell 37*, 1027–1033.
Pai, E. F., Kabsch, W., Krengel, U., Holmes, K., John, J., & Wittinghofer, A. (1989) *Nature 341*, 209–214.
Pai, E. F., Krengel, U., Petsko, G. A., Goody, R. S., Kabsch, W., & Wittinghofer, A. (1990) *EMBO J. 9*, 2351–2359.
Palme, K., Diefenthal, T., Sander, C., Vingron, M., & Schell, J. (1989) in *The guanine-nucleotide binding proteins* (Bosch, L., Kraal, B., & Parmeggiani, A., Eds.) pp 273–284, Plenum, New York.
Pizon, V., Lerosey, I., Chardin, P., & Tavitian, A. (1988a) *Nucleic Acids Res. 16*, 7719.
Pizon, V., Chardin, P., Lerosey, I., Olofsson, I., & Tavitian, A. (1988b) *Oncogene 3*, 201–204.
Polakis, P. G., Snyderman, R., & Evans, T. (1989) *Biochem. Biophys. Res. Commun. 160*, 25–32.
Powers, S., Kataoka, T., Fasano, O., Goldfarb, M., Strathem, J., Broach, J., & Wigler, M. (1984) *Cell 36*, 607–612.
Reinstein, J., Schlichting, I., & Wittinghofer, A. (1990) *Biochemistry 29*, 7451–7459.
Reymond, C. D., Gomer, R. H., Mehdy, M. C., & Firtel, R. A. (1984) *Cell 39*, 141–148.
Robbins, S. M., Williams, J. G., Jermyn, K. A., Spiegelman, G. B., & Weeks, G. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 938–942.
Salminen, A., & Novick, P. (1987) *Cell 49*, 527–538.
Sander, C., & Schneider, R. (1991) *Proteins: Struct., Funct., Genet. 9*, 56–68.
Saxe, S. A., & Kimmel, A. R. (1988) *Dev. Genet. 9*, 259–265.
Schafer, W. R., Kim, R., Sterne, R., Thorner, J., Kim, S. H., & Rine, J. (1989) *Science 245*, 379–385.
Schejter, E., & Shilo, B.-Z. (1985) *EMBO J. 4*, 407–412.
Schlichting, I., Almo, S. C., Rapp, G., Wilson, K., Petratos, K., Lentfer, A., Wittinghofer, A., Kabsch, W., Pai, E. F.,

Petsko, G. A., & Goody, R. S. (1990) *Nature 345*, 309-314.
Seeburg, P. H., Colby, W. W., Capon, D. J., Goeddel, D. V., & Levinson, A. D. (1984) *Nature 312*, 71-77.
Sigal, I. S., Gibbs, J. B., D'Alonzo, J. S., & Scolnick, E. M. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 4725-4729.
Smith, T. F., & Waterman, M. S. (1981) *J. Mol. Biol. 147*, 195-197.
Spandidos, D. A., & Anderson, M. L. (1989) *J. Pathol. 157*, 1-10.
Swanson, M., Elste, A., Greenberg, S., Schwartz, J., Aldrich, T., & Furth, M. (1986) *J. Cell Biol. 103*, 485-492.
Taparowsky, E., Shimizu, K., Goldfarb, M., & Wigler, M. (1983) *Cell 34*, 581-586.
Touchot, N., Chardin, P., & Tavitian, A. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 8210-8214.
Trahey, M., & McCormick, F. (1987) *Science 238*, 542-545.
Valencia, A., Kjeldgaard, M., Pai, E. F., & Sander, C. (1991) *Proc. Natl. Acad. Sci. U.S.A.* (in press).
Vielh, E., Touchot, N., Zahraoui, A., & Tavitian, A. (1989) *Nucleic Acids Res. 17*, 1770.
Westaway, D., Papkoff, J., Moscovici, C., & Varmus, H. E. (1986) *EMBO J. 5*, 301-309.
Wierenga, R. K., Terpstra, P., & Hol, W. G. J. (1986) *J. Mol. Biol. 187*, 101-107.
Yeramian, P., Chardin, P., Madaule, P., & Tavitian, A. (1987) *Nucleic Acids Res. 4*, 1869.
Zahraoui, A., Touchot, N., Chardin, P., & Tavitian, A. (1988) *Nucleic Acids Res. 16*, 1204.
Zahraoui, A., Touchot, N., Chardin, P., & Tavitian, A. (1989) *J. Biol. Chem. 264*, 12394-12401.

Accelerated Publications

Primary Donor Structure and Interactions in Bacterial Reaction Centers from Near-Infrared Fourier Transform Resonance Raman Spectroscopy†

Tony A. Mattioli,*,‡ Andreas Hoffmann,§ Bruno Robert,‡ Bernhard Schrader,§ and Marc Lutz‡

Département de Biologie Cellulaire et Moléculaire, Centre d'Etudes de Saclay, 91191 Gif-sur-Yvette Cedex, France, and Institut für Physikalische und Theoretische Chemie, Universität Essen, 4300 Essen 1, Germany

Received January 28, 1991; Revised Manuscript Received March 18, 1991

ABSTRACT: Preresonance Raman and resonance Raman spectra of the primary donor (P) from reaction centers of the *Rhodobacter* (*Rb.*) *sphaeroides* R26 carotenoidless strain in the P and P^+ states, respectively, were obtained at room temperature with 1064-nm excitation and a Fourier transform spectrometer. These spectra clearly indicate that the chromophore modes are observable over those of the protein with no signs of interference below 1800 cm^{-1}. The chromophore modes are dominated by those of the bacteriochlorophylls (BChl *a*), and it is estimated that, in the P state, ca. 65% of the Raman intensity of the BChl *a* modes arises from the primary donor. This permits the direct observation of a vibrational spectrum of the primary donor at preresonance with the excitonic 865-nm band. The Raman spectrum of oxidized reaction centers in the presence of ferricyanide clearly exhibits bands arising from a BChl a^+ species. The magnitude of the frequency shift of a keto carbonyl of neutral P from 1691 to 1717 cm^{-1} upon P^+ formation strongly suggests that one BChl molecule in P^+ carries nearly the full +1 charge. Our results indicate that the unpaired electron in $P^{\bullet+}$ does not share a molecular orbital common to the two components of the dimer on the time scale of the resonance Raman effect (ca. 10^{-13} s).

The primary events in bacterial photosynthesis occur in membrane-bound proteins known as reaction centers (RCs).[1] The isolated RC consists of six bacteriochlorin pigments (four bacteriochlorophyll *a* and two bacteriopheophytin *a* molecules), two quinones, one non-heme iron, one carotenoid molecule, and approximately 850 amino acid residues contained in three polypeptide subunits named L, M, and H. Within the RC, electron transfer originates from the primary donor P, which consists of a pair of bacteriochlorophyll (BChl) molecules in mutual excitonic interaction. Although the X-ray crystallographic structures of the RC from *Rhodopseudomonas* (*Rps.*) *viridis* (Deisenhofer & Michel, 1989) and *Rhodobacter* (*Rb.*) *sphaeroides* (Allen et al., 1987a,b; Chang et al., 1986; Tiede et al., 1988) are resolved, the understanding of charge separation and stabilization requires a thorough characterization of the physicochemical properties of P and its cation radical, $P^{\bullet+}$.

The absorption spectrum of bacterial reaction centers exhibits a broad band in the near-infrared that corresponds to the first excited singlet state of the primary donor pair, ^1P. For bacteriochlorophyll *a* (BChl *a*) containing RCs, such as *Rb. sphaeroides*, this band appears at ca. 870 nm. The characterization of this band in an attempt to explain the asymmetric functioning of the RC has been the subject of recent intensive work [for a review, see Friesner and Won (1989)]. When P undergoes one-electron chemical or pho-

† T.A.M. gratefully acknowledges fellowships from NATO/NSERC (Canada) and EMBO.
* Author to whom correspondence should be addressed.
‡ CE Saclay.
§ Universität Essen.

[1] Abbreviations: RR, resonance Raman; NIR, near-infrared; FT, Fourier transform; RC, reaction center; *Rb.*, *Rhodobacter*; *Rps.*, *Rhodopseudomonas*; *Rsp.*, *Rhodospirillum*; BChl, bacteriochlorophyll; BPhe, bacteriopheophytin; EPR, electron paramagnetic resonance; THF, tetrahydrofuran.

0006-2960/91/0430-4648$02.50/0