PROTEINS: Structure, Function, and Genetics 17:363–374 (1993)

Hundreds of Ankyrin-Like Repeats in Functionally Diverse Proteins: Mobile Modules That Cross Phyla Horizontally?

Peer Bork
Max-Delbrück-Centre of Molecular Medicine, D-13125 Berlin, Germany

ABSTRACT **Based on pattern searches and systematic database screening, almost 650 different ankyrin-like (ANK) repeats from nearly all phyla have been identified; more than 150 of them are reported here for the first time. Their presence in functionally diverse proteins such as enzymes, toxins, and transcription factors strongly suggests domain shuffling, but their occurrence in prokaryotes and yeast excludes exon shuffling. The spreading mechanism remains unknown, but in at least three cases horizontal gene transfer appears to be involved. ANK repeats occur in at least four consecutive copies. The terminal repeats are more variable in sequence. One feature of the internal repeats is a predicted central hydrophobic α-helix, which is likely to interact with other repeats. The functions of the ankyrin-like repeats are compatible with a role in protein–protein interactions.** © 1993 Wiley-Liss, Inc.

Key words: **sequence analysis, homology search, ANK repeat, horizontal gene transfer, cell cycle proteins, transcription factor NF-κB**

INTRODUCTION

In order to trace protein evolution, it is advisable to follow the distribution of protein domains, i.e., structurally and functionally independent building blocks (modules). Whereas the use of homologous domains seems to be rather limited in phylogenetically "old" enzymes (typical examples are the dinucleotide-binding domains of oxidoreductases), many "modern" proteins, such as most extracellular animal proteins, consist of a variety of domains.[1,2a,b] These modules appear to be present in functionally diverse proteins and are thought to be the result of exon shuffling.[3] Further mechanisms by which to spread protein domains through the genome are suggested by prokaryotic systems which contain shuffled domains such as glycohydrolases,[4] "two-component" signal transduction proteins,[5] and phosphoenolpyruvate: sugar phosphotransferases.[6] There might even be horizontal gene transfer involved in domain shuffling as has recently been proposed for bacterial fibronectin type III modules which have apparently been acquired from animals.[7]

To continue the studies on the structure, function, and evolution of protein domains that are widespread among different phyla, I examined another module which apparently does not need "exon shuffling" to cover a wide range of organisms: the ankyrin-like (ANK) repeat.[8] Ankyrins are proteins that are believed to couple a variety of integral membrane proteins to spectrin.[9] Although they contain 24 ANK copies, the protein–protein interactions have been assigned to particular repeats, e.g., repeats 21–22 have been shown to be responsible for high affinity binding of the anion exchanger (reviewed in ref. 9).

First discovered as homologous regions between some cell cycle proteins ("CDC10/SW14 repeat") and the *Drosophila* protein *notch,*[10] the ANK- repeat has subsequently been detected in various regulatory proteins (summarized in refs. 9, 11a). The repeat, which has a length of about 33 amino acids, has also been noted in several poxviruses[8,12] and in mouse mammary tumor virus.[13] In the latter case, a part of a *notch*-like protein has apparently been very recently incorporated into the virus genome. Not only the location in both extracellular and intracellular proteins is noteworthy, but also the occurrence in functionally diverse proteins of different phyla; apart from animals and yeast, several ANK repeats have recently been detected in a plant protein[14] and even in prokaryotes[11a] (A. Neuwald, personal communication; this work).

Here, the results of an extensive screening of current sequence databases are presented which include the identification of numerous additional ANK repeats. Based on these data, a comprehensive analysis of all recognized ANK repeats has been carried out in order to obtain information about the

Abbreviations: ANK, ankyrin-like repeat; NF-κB, nuclear factor κB; TNF, tumor necrosis factor.

Received July 1, 1993; revision accepted August 12, 1993.

Address reprint requests to Dr. Peer Bork at his present address: European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69012 Heidelberg, Germany.

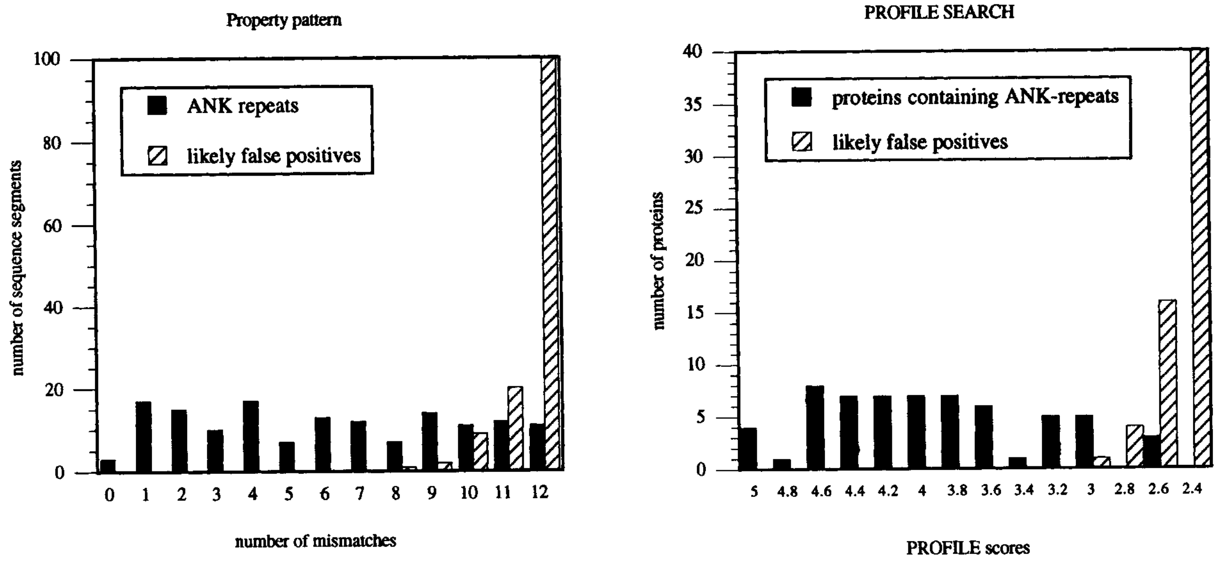


Fig. 1. Statistics of pattern database searches with two fused ANK repeats. The separation from the random background of nonrelated sequences in SWISSPROT and PIR is shown for both consensus property patterns[18] and profiles.[19] The number of hits detected by the property patterns is higher because it is able to recognize multiple copies within a single protein. PROFILESEARCH detects only the best-scoring repeat within one protein. Note that the real number of ANK repeats is much higher. Often they are present in nucleic acid databases but not yet translated or not recognized as open reading frames.

structural, functional, and evolutionary constraints for this extremely widespread protein module.

METHODS

All sequences containing ANK repeats have been subjected to a number of sequence analysis methods (for details see ref. 15). The screening scheme can be summarized by the following steps: (1) Known ankyrin-like repeats were extracted from SWISS-PROT[16] if their presence was indicated in this protein sequence database. The phasing of the repeats proposed by Michaely and Bennett[11a] has been used because this is consistent with exon borders in ankyrin and with the requirement for complete terminal repeats in numerous proteins. (2) A multiple alignment using PILEUP[17] of all indicated ANK repeats was then performed in order to (3) construct consensus patterns[18] and PROFILES.[19] (4) With these constructs several sequence databases were screened (SWISS-PROT, release 25; PIR release 26; EMBL, release 34). (5) The search was conducted in three iterations of adding clearly identified members to the learning set (multiple alignment) and recalculating the consensus patterns. A fourth iteration did not reveal any additional member of the family. Although no routine has yet been implemented for automatically stopping the iteration procedure, in all the cases tested so far,[18] three to four iterations have led to a sensitive pattern. When false positives are randomly included in the learning set, the pattern's no longer able to discriminate members of the learning set from the random background of unrelated sequences (noise). The noise can be estimated by considering the best scores of definite false positives, e.g., proteins with known 3D structure or other well-characterized proteins. Since ANK repeats have been exclusively found in consecutive copies, the separation of ANK repeats from the random background of unrelated sequences could be improved by fusing two repeats for database searches (Fig. 1). Proteins were considered to contain ANK repeats if they were detected either by the PROFILES[19] or by the property patterns[18] above the random background of nonrelated sequences (Fig. 1). (6) As an additional check, all proteins found to contain ANK repeats were subjected to (T)FASTA homology searches[20] against different sequence databases. (7) In a final round, several candidates with weak signals were inspected in detail. Current scoring schemes of homology search methods are very much dependent on empirical parameters such as the substitution matrices used or amino acid composition of the database. Although the use of family information certainly helps to justify subtle similarities, a final judgment for low scoring sequences is often context-dependent and requires a combination of methods. Therefore, a few putative ANK repeats not detectable by automatic methods were added manually (marked in Fig. 4) when they have a weak signal and (a) they occur in between clearly identified repeats or (b) they correspond to less conserved terminal repeats in proteins where less than 6 repeats were clearly identified. For these repeats, no significance criterion can be given, but considering their location in between or next to other ANK-repeats, it is likely that they represent rather divergent copies. (8) Programs of the GCG package[17] were used for sequence clustering and construction of dendrograms.

In vaccinia virus (strain Copenhagen[21]) all proteins located next to the putative ANK repeats on the genome were studied in detail in order to find functional correlations between ANK repeats and neighboring proteins.

RESULTS

Database Screening

With the methods used, the number of classified ANK repeats in current sequence databases was dramatically increased by 165 to 639 (not counting nearly identical sequences of one species). Even if highly similar sequences (> 75% amino acid identity) and putative orthologues (i.e., proteins encoded by equivalent genes in different species) are excluded, about 250 eukaryotic ANK repeats were found (Figs. 1 and 2). Due to the higher mutation rate in viruses, most of the about 240 repeats in several poxviruses also met this criterion (< 75% amino acid identity to any other repeat). Since the complete genome of vaccinia virus (strain Copenhagen[21]) is stored in current databases, only vaccinia virus proteins have been analyzed further (Fig. 3). Most of these proteins (shown in Fig. 3) have counterparts in related poxviruses, although some of the copies might have been lost during the course of virus evolution.[22]

The applied searching scheme allowed both the identification of many additional copies in proteins for which several ANK repeats have already been reported and the recognition of ANK repeats in proteins with no known similarities (Figs. 2 and 3). Examples for the latter group are a putative protein of yeast chromosome III (YCR51W) and *Pho81*, another yeast protein believed to regulate the *Ph1* activator.[23] Furthermore, the C-terminus of an unidentified open reading frame next to *Pho81,* but on the opposite strand could be also shown to be largely composed of ANK repeats (*Pho82* in Figs. 2 and 4). A (hypothetical) protein with 7 ANK repeats was found in *E. coli* and even an RNase[24] could be clearly identified to contain 9 ANK repeats (Figs. 2 and 4).

Homology searches in nucleic acid databases often reveal sequencing errors,[25] e.g., if the query protein matches two consecutive regions of two shifted reading frames. The analysis of another prokaryotic protein identified to containing ANK repeats, *Phlb* from *Serratia liquefaciens,*[26] revealed such a putative frameshift; the correction of which would extend the number of ANK repeats from 4 to 6 (Fig. 5).

ANK Repeats as Consecutive Copies

One interesting feature of the ANK repeats is the occurrence of at least 4 consecutive copies per protein. The analysis presented here revealed additional copies in all proteins for which only one or two repeats had been reported previously. Examples are the *Drosophila* calmodulin-binding protein *trpl*[27] and its relative *trp,* a phototransduction gene product.[28] For both proteins, three more remote ANK repeats could be added to the one reported copy[27]; together they cover a substantial part of the N-terminal domain (Figs. 2 and 4). A similar situation is found in the yeast cell-cycle proteins *SWI4, SWI6, Res1,* and *cdc10,* in which two nonconsecutive copies have been reported.[10,29] The pattern searches identify two additional copies which completely cover the segment between the known ANK repeats (Figs. 2 and 4). This newly defined domain is of functional importance, because SWI4/SWI6 and probably res1/cdc10 form transcription factor complexes[29] and mutations in the ANK repeats reduce DNA binding.[30,31]

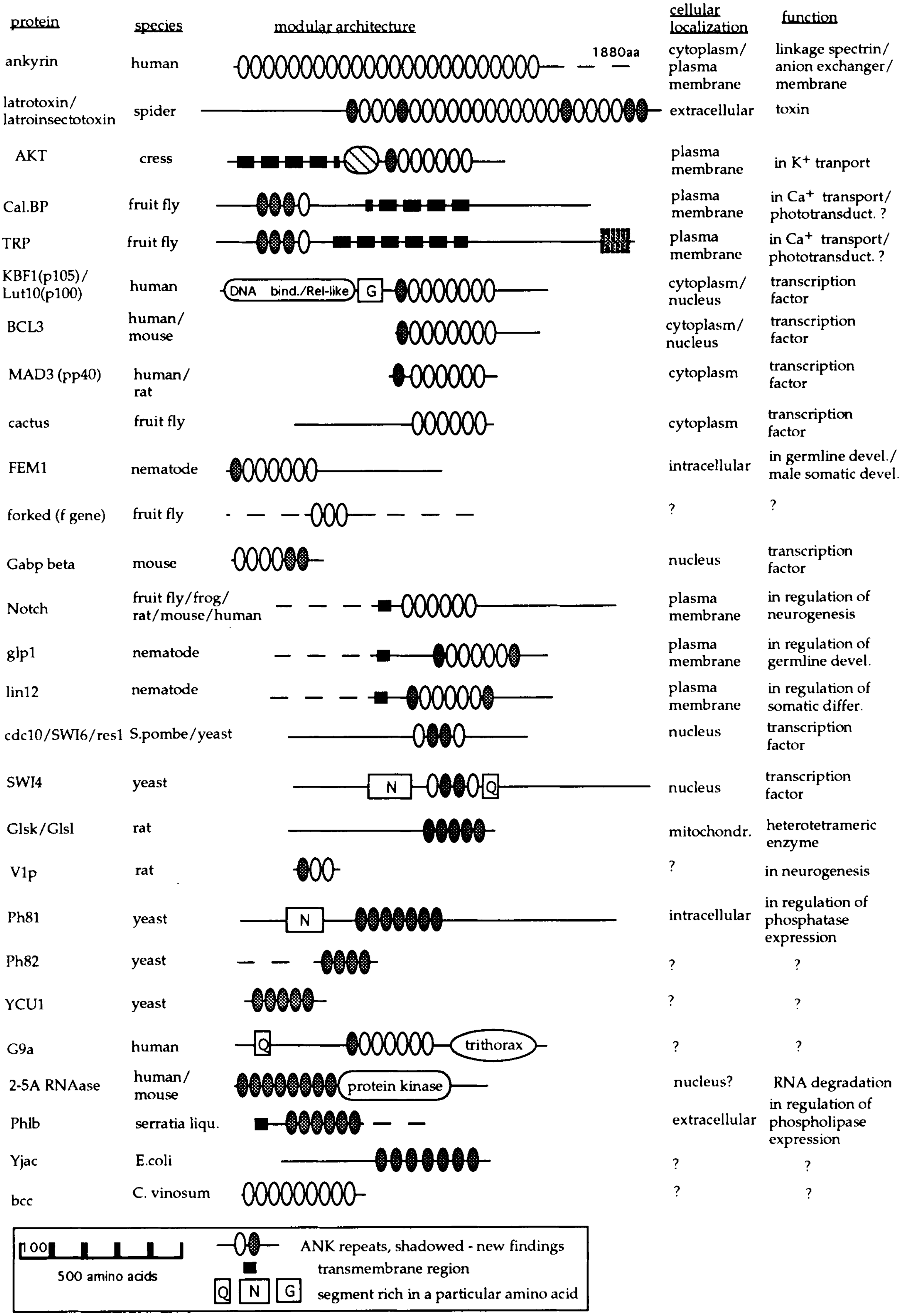
Only one eukaryotic protein, the rat cerebellar protein V-1, does not fit into this scheme, as it consists only of three consecutive ANK-repeats (Figs. 2 and 4). It remains to be studied whether the isolated protein[31] has arisen from a larger functional active precursor.

Although the majority of poxvirus proteins consist of numerous consecutive ANK repeats, there are a few short putative proteins which contain only one ANK repeat (Fig. 2). Comparison of the related vaccinia and variola virus genomes[22] has revealed the loss of ANK repeats in corresponding proteins. Even in very closely related strains of vaccinia virus (e.g., strain WR compared to Copenhagen) stop codons have been inserted leading to fragmentary proteins.[32] The functionality of the resulting short open reading frames with less than 4 ANK repeats remains to be proven.

Conservation of ANK Repeats

With the increasing number of ANK repeats added to current sequence databases, it becomes obvious that many copies deviate from the original consensus sequence.[8] The existence of much more divergent repeats can not be excluded, e.g., the methods used detected a weak signal of a fifth (N-terminal) repeat in *SWI4, SWI6, cdc10,* and *res1* and a seventh repeat in *notch,* but they could not be verified by pairwise comparisons and multiple alignment techniques and were therefore not included.

Although the length of the repeat is in many cases 33 amino acids, large insertions of up to 13 amino acids occur.[12] Insertions frequently occur in the virus proteins and mainly at position 15 (Fig. 4). Thus, in both property patterns and profiles, large gaps were allowed at this position. Not a single position of the alignment (Fig. 4) contains an invariant amino acid. However, several strictly conserved hydrophobic positions can be observed from the alignment of over 300 repeats (Figs. 4 and 6). They are flanked by polar or hydrophilic positions (Figs. 4 and 6) result-



protein	species	modular architecture	cellular localization	function
ankyrin	human	1880aa	cytoplasm/ plasma membrane	linkage spectrin/ anion exchanger/ membrane
latrotoxin/ latroinsectotoxin	spider		extracellular	toxin
AKT	cress		plasma membrane	in K$^+$ tranport
Cal.BP	fruit fly		plasma membrane	in Ca$^+$ transport/ phototransduct. ?
TRP	fruit fly		plasma membrane	in Ca$^+$ transport/ phototransduct. ?
KBF1(p105)/ Lut10(p100)	human	DNA bind./Rel-like G	cytoplasm/ nucleus	transcription factor
BCL3	human/ mouse		cytoplasm/ nucleus	transcription factor
MAD3 (pp40)	human/ rat		cytoplasm	transcription factor
cactus	fruit fly		cytoplasm	transcription factor
FEM1	nematode		intracellular	in germline devel./ male somatic devel.
forked (f gene)	fruit fly		?	?
Gabp beta	mouse		nucleus	transcription factor
Notch	fruit fly/frog/ rat/mouse/human		plasma membrane	in regulation of neurogenesis
glp1	nematode		plasma membrane	in regulation of germline devel.
lin12	nematode		plasma membrane	in regulation of somatic differ.
cdc10/SWI6/res1	S.pombe/yeast		nucleus	transcription factor
SWI4	yeast	N Q	nucleus	transcription factor
Glsk/Glsl	rat		mitochondr.	heterotetrameric enzyme
V1p	rat		?	in neurogenesis
Ph81	yeast	N	intracellular	in regulation of phosphatase expression
Ph82	yeast		?	?
YCU1	yeast		?	?
G9a	human	Q trithorax	?	?
2-5A RNAase	human/ mouse	protein kinase	nucleus?	RNA degradation
Phlb	serratia liqu.		extracellular	in regulation of phospholipase expression
Yjac	E.coli		?	?
bcc	C. vinosum		?	?

100
500 amino acids
ANK repeats, shadowed - new findings
transmembrane region
Q N G
segment rich in a particular amino acid

Fig. 2.

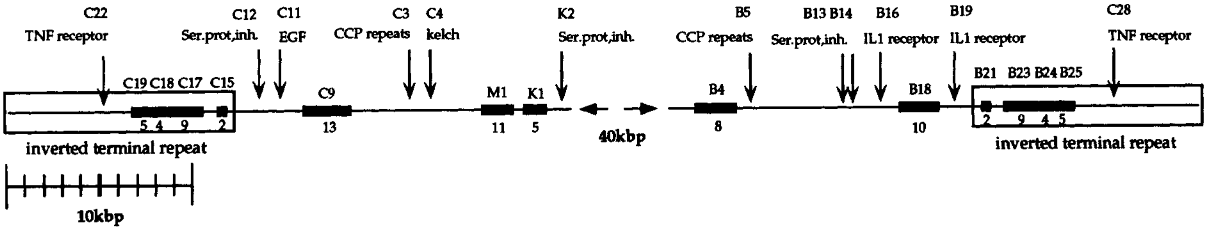


Fig. 3. Location of ANK repeats (thick lines) in vaccinia virus (Copenhagen strain) for which the whole genome (nearly 192 kbp) has been sequenced.[21] Other poxviruses contain homologous sequences. The identified ANK repeats (copy numbers are given below the corresponding protein names) are only found near or within the inverted terminal repeats. The central part of the virus (not shown) mainly contains essential genes necessary for replication.[21] In addition to the ANK repeat containing proteins, the location of extracellular proteins is shown (CCP, complement control repeat or sushi repeat; kelch, regulatory egg chamber protein).[49] Only two protein kinases (B1, B12) and relatives of T2, T3, T7, T8 in Shope fibroma virus have been mapped to the displayed vaccinia virus segments. Some of the proteins consisting of ANK repeats could only be identified because of their overall homology to other vaccinia proteins that contain more conserved ANK copies, e.g., C15/B21 and C19/B25 are similar to the N-terminus of B18 and C18/B24 are similar to the C-terminus of B4. The 19 previously identified ANK repeats in vaccinia virus (compare with Fig. 4b) have recently been summarized by Shchelkunov et al.[22] Note the symmetric location of the ANK repeat containing proteins within the virus which might reflect an ancient inversion of a large part of the viral genome.

ing in a characteristic consensus property pattern[18] for this domain. Spacer regions in between repeats have been observed. They do not exceed 20 amino acids.

Secondary structure predictions using the profile based neural network method PHD[33] suggest the presence of an α-helix in the central part of the repeat (positions 16–25) with a turn at either end (Figs. 4 and 6). An additional turn is predicted within the five N-terminal residues of the repeats. The two remaining regions (positions 6–12 and 28–33) could not be clearly assigned, although they might form secondary structure elements. Two different models for the tertiary structure and the packing of the consecutive repeats have been proposed,[11,34] assuming an N-terminal α-helix.[8] Mapping conserved positions onto a helix wheel (Fig. 6) indicates a nearly buried central helix. Since the repeats are only 33 residues long, several tightly packed ANK repeats might interact. This can happen in a circular, barrel-like arrangement as found in the packing of consecutive repeats in the propeller structures of influenza neuramidase, methylamine dehydrogenase, galactose oxidase,[35] and regulatory kelch repeats (Bork and Doolittle, unpublished). On the other hand, a tight linear arrangement of several ANK copies might be responsible for the conserved hydrophobic N-terminal and central positions (Fig. 4).

Linear Arrangement of ANK Repeats

When evaluating sequence conservation, it becomes obvious that the terminal repeats have many more deviations from the general consensus than those located centrally (Table I). This is supported by multiple alignments of different sets of overall related sequences such as ankyrins, spider toxins, the nuclear factor κB (NF-κB) family, or *notch* orthologues. These comparisons also reveal a lower conservation of the external repeats (data not shown). Because of this sequence variability, numerous external repeats have not yet been detected in sequence databases or have only been reported as "partial" repeats (see Fig. 2). Taking into account (1) the short length of the repeat with a maximum of three secondary structure elements; (2) the high hydrophobicity of the internal repeats relative to their physical size; (3) the sequence variability of the terminal repeats; and (4) the occurrence in consecutive copies, a tight packing arrangement is suggested in which the repeats are stacked on top of each other. Both the central α-helix (Fig. 6) and a conserved hydrophobic patch (positions 6–10 in Fig. 4) have to interact with other repeats to be shielded from the surface. Nevertheless, some of the hydrophobic positions might point outward to form sites for protein–protein interactions.

ANK Repeats and Protein–Protein Interactions

The occurrence of ANK repeats in extracellular proteins such as spider toxins[36] and the phospholipase regulator *Phlb*,[26] mitochondrial enzymes such as glutaminases,[37] nuclear cell cycle regulators (*cdc10, SWI4, SWI6, Res1*[30]), cytoplasmic proteins

Fig. 2. Distribution of ANK repeats in eukaryotes and prokaryotes. The length of the lines corresponds to the length of the proteins (counted in amino acids). Stippled lines indicate that only fragments have been sequenced or that certain parts of the proteins have been omitted. Boxes and ellipsoides represent structural domains including the ANK repeats. Although the average length of an ANK repeat is 33 amino acids, deviations such as the insertion of 13 amino acids are not unusual. Note the functional variety and the different cellular locations. Most of the proteins are present in SWISSPROT or PIR protein databases; others have been stored in nucleic acid databases and can be found via the following EMBL accession numbers: *cactus* (L04964), *akt1* (X62907), *ph82* (X52482), *clpb* (M88185), lat2 (Z14086), bcc (L13414), 2-5A RNase (L10382), G9a (X69838). Forked (F gene; X69871) is located on the X chromosome but has not yet been described in detail.

```
pred.sec.struc.:                 aaaaaaaaa
consensus:             t otLHhAh  tt thht LLt t t
 Akt1/Artha-1*    484 IMNNLLQHLKEMNDPVMTNVLLEIENMLARGKL
 Akt1/Artha-2     517 DLPLNLCFAAIREDDLLLHQLLKRGLDPNESDN
 Akt1/Artha-3     550 NGRTPLHIAASKGTLNCVLLLLEYHADPDDITL
 Akt1/Artha-4     583 EGSVPLWEAMVEGHEKVVKVLLEHGSTIDAGDV
 Akt1/Artha-5     614 DVGHFACTAAEQGNLKLLKEIVLHGGDVTRPRR
 Akt1/Artha-6     647 TGTSALHTAVCEENIEMVKYLLEQGADVNKQDM
 Akt1/Artha-7     689 HGWTPRDLAEQQGH.EDIKALFREKLHERRVHI
 Ank1_Human-1      10 DAATSFLRAARSGNLDKALDHLRNGVDINTCNQ
 Ank1_Human-2      43 NGLNGLHLASKEGHVKMVVELLHKEIILETTTK
 Ank1_Human-3      76 KGNTALHIAALAGQDEVVRELVNYGANVNAQSQ
 Ank1_Human-4     109 KGFTPLYMAAQENHLEVVKFLLENGANQNVATE
 Ank1_Human-5     142 DGFTPLAVALQQGH.ENVAHLINYGTKGKVTEQ
 Ank1_Human-6     171 VRLPALHIAARNDDTRTAAVLLQNDPNPDVLSK
 Ank1_Human-7     204 TGFTPLHIAAHYENLNVAQLLLNRGASVNFTPQ
 Ank1_Human-8     237 NGITPLHIASRRGNVIMVRLLLDRGAQIETKTK
 Ank1_Human-9     270 DELTPLHCAARNGHVRISEILLDHGAPIQAKTK
Ank1_Human-10     303 NGLSPIHMAAQGDHLDCVRLLLQYDAEIDDITL
Ank1_Human-11     336 DHLTPLHVAAHCGHHRVAKVLLDKGAKPNSRAL
Ank1_Human-12     369 NGFTPLHIACKKNHVRVMELLLKTGASIDAVTE
Ank1_Human-13     402 SGLTPLHVASFMGHLPIVKNLLQRGASPNVSNV
Ank1_Human-14     435 KVETPLHMAARAGHTEVAKYLLQNKAKVNAKAK
Ank1_Human-15     468 DDQTPLHCAARIGHTNMVKLLLENNANPNLATT
Ank1_Human-16     501 AGHTPLHIAAREGHVETVLALLEKEASQACMTK
Ank1_Human-17     534 KGFTPLHVAAKYGKVRVAELLLERDAHPNAAGK
Ank1_Human-18     567 NGLTPLHVAVHHNNLDIVKLLLPRGGSPHSPAW
Ank1_Human-19     600 NGYTPLHIAAKQNQVEVARSLLQYGGSANAESV
Ank1_Human-20     633 QGVTPLHLAAQEGHAEMVALLLSKQANGNLGNK
Ank1_Human-21     666 SGLTPLHLVAQEGHVPVADVLIKHGVMVDATTR
Ank1_Human-22     699 MGYTPLHVASHYGNIKLVKFLLQHQADVNAKTK
Ank1_Human-23     732 LGYSPLHQAAQQGHTDIVTLLLKNGASPNEVSS
Ank1_Human-24     765 DGTTPLAIAKRLGY4DVLKVVTDETSFVLVSDK
 Bcl3_Human-1*     60 AVPGPPHGLARPEALYYPGALLPLYPTRAMGSP
 Bcl3_Human-2     126 DGDTPLHIAVVQGNLPAVHRLVNLFQQGGRELD
 Bcl3_Human-3     163 LRQTPLHLAVITTLPSVVRLLVTAGASPMALDR
 Bcl3_Human-4     196 HGQTAAHLACEHRSPTCLRALLDSAAPGTLDLE
 Bcl3_Human-5     233 DGLTALHVAVNTECQETVQLLLERGADIDAVDI
 Bcl3_Human-6     267 SGRSPLIHAVENNSLSMVQLLLQHGANVNAQMY
 Bcl3_Human-7     300 SGSSALHSASGRGLLPLVRTLVRSGADSSLKNC
 Bcl3_Human-8     333 HNDTPLMVARSRRVIDILRGKATRPASTSQPDP
 Cc10_Schpo-1     356 LGHAALHWAAAVAKMPLLQALIHKGANPLRGNL
 Cc10_Schpo-2*    389 TGETALMRSVLVTN4NSFGDLLDLLYASLPCTD
 Cc10_Schpo-3*?   426 AGRTVVHHICLTAG9YYLETLLNWAKKHASGNN
 Cc10_Schpo-4     483 NGDTALNIAARIGNKNIVEVLMQAGASAYIPNR
 Fem1_Caeel-1*      8 FRTVIYNAAAVGNLQRIKVFTINSRNDRQWIID
 Fem1_Caeel-2      47 DGRYPLVIAARNGHANVVEYLLEIGADPSVRGV
 Fem1_Caeel-3      88 QGTPPLWAASAAGHIEIVKLLIEKANADVNQAT
 Fem1_Caeel-4     122 TRSTPLRGACYDGHLDIVKYLLEKGADPHIPNR
 Fem1_Caeel-5     155 HGHTCLMIASYRNKVGIVEELLKTGIDVNKKTE
 Fem1_Caeel-6     188 RGNTALHDAAESGNVEVVKILLKHGSVLMKDIQ
 Fem1_Caeel-7     220 QGVDPLMGAALSGFLDVLNVLADQMPSGIHKRD
 Gab1_Mouse-1       5 DLGKKLLEAARAAQDDEVRILMANGAPFTTDWL
 Gab1_Mouse-2      37 LGTSPLHLAAQYGHFSTTEVLLRAGVSRDARTK
 Gab1_Mouse-3      70 VDRTPLHMAASEGHANIVEVLLKHGADVNAKDM
 Gab1_Mouse-4     103 LKMTALHWATEHNHQEVVELLIKYGADVHTQSK
 Gab1_Mouse-5*    136 FCKTAFDISIDNGN4EILQIAMQNQINTNPESP
 Gab1_Mouse-6*    171 PDTVTIHAATPQFI3GGVVNLTDETGVSAVQFG
 Glp1_Caeel-1*    890 ADEIPLHVQAAGPD.AITAPITNESVNQVDSKY
 Glp1_Caeel-2     921 YRRRVLHWLAANVR8TEAIRCLKAGADVNARDC
 Glp1_Caeel-3     961 DENTALMLAVRAHRVRLSVVLLREGANPTIFNN
 Glp1_Caeel-4     994 SERSALHEAVVNKDLRILRHLLTDKRLLKEIDE
 Glp1_Caeel-5    1030 NGMTALMLVARELG4EMAELLLSKGAKLDYDGA
 Glp1_Caeel-6    1074 KGRTALHYAAMHDNEEMVIMLVRRSSNKDKQDE
 Glp1_Caeel-7    1107 DGRTPIMLAAKEGCEKTVQYLALNDASLGIVDS
 Glp1_Caeel-8*?1139 SMDMTAAQVAEASYHHELAAFLRQVANERHRND
   Glsk_Rat-1*    474 SGQFAFHVGLPAKSGVAGGILLVVPNVMGMMCW
   Glsk_Rat-2*    515 NSVKGIHFCHDLVS7DNLRHFAKKLDPRREGGD
   Glsk_Rat-3*    558 KSVINLLFAAYTGDVSALRRFALSAMDMEQRDY
   Glsk_Rat-4*    590 DSRTALHVAAAEGHVEVVKFLLEACKVNPFPKD
   Glsk_Rat-5*?   624 WNNTPMDEALHFGH.HDVFKILQEYQVQYTPQG
 Kbf1_Human-1*    509 LAKRHANALFDYAVTGDVKMLLAVQRHLTAVQD
 Kbf1_Human-2     543 NGDSVLHLAIIHLHSQLVRDLLEVTSGLISDDI
 Kbf1_Human-3     582 LYQTPLHLAVITKQEDVVEDLLRAGADLSLLDR
 Kbf1_Human-4     615 LGNSVLHLAAKEGHDKVLSILLKHKKAALLLDH
 Kbf1_Human-5     651 DGLNAIHLAMMSNSLPCLLLLVAAGADVNAQEQ
 Kbf1_Human-6     685 SGRTALHLAVEHDNISLAGCLLLEGDAHVDSTT
 Kbf1_Human-7     719 DGTTPLHIAAGRGSTRLAALLKAAGADPLVENF
 Kbf1_Human-8     772 PGTTPLDMATSWQVFDILNGKPYEPEFTSDDLL
 Lata_Latma-1*    453 DIDRDLYNAASNPD2VGFKEFTKLNYDGANIRA
 Lata_Latma-2     490 HGRTVFHAAAKSGNDKIMFGLTFLAKSTELNQP
 Lata_Latma-3     525 KGYTPIHVAADSGNAGIVNLLIQRGVSINSKTY
 Lata_Latma-4     559 FLQTPLHLAAQRGFVTTFQRLMESPEININERD
 Lata_Latma-5*    593 DGFTPLHYAIRGGE.RILEAFLNQISIDVNAKS
 Lata_Latma-6     626 TGLTPFHLAIIKNDWPVASTLLGSKKVDINAVD
 Lata_Latma-7     660 NNITALHYAAILGYLETTKQLINLKEINANVVS
 Lata_Latma-8     695 GLLSALHYAILYKHDDVASFLMRSSNVNVNLKA
 Lata_Latma-9     729 GGITPLHLAVIQGRKQILSLMFDIGVNIEQKTD
Lata_Latma-10     762 GKYTPLHLAAMSKYPELIQILLDQGSNFEAKTN
Lata_Latma-11     795 SGATPLHLATFKGKSQAALILLNNEVNWRDTDE
Lata_Latma-12     828 NGQMPIHGAAMTGLLDVAQAIISIDATVVDIED
Lata_Latma-13     862 NSDTPLNLAAQNSHIDVIKYFIDQGADINTRNK
Lata_Latma-14     895 KGLAPLLAFSKKGNLDMVKYLFDKNANVYIADN
Lata_Latma-15     928 DGMNFFYYAVQNGHLNIVKYAMSEKDKFEWSNT
Lata_Latma-16     975 ISHFAVCDAVQFDRIEIVKYFVGTLGNFAICGP
Lata_Latma-17    1003 AICGPLHQAARYGHLDIVKYLVEEEFLSVDGSK
Lata_Latma-18*   1035 KTDTPLCYASENGHFTVVQYLVSNGAKVNHDCG
Lata_Latma-19    1068 NGMTAIDKAITKNHLQVVQFLAANGVDFRRKNS
Lata_Latma-20    1101 RGTTPFLTAVAENALDIAEYLIREKRQDININE
Lata_Latma-21    1137 DKDTALHLAVYYKNLQMIKLLIKYGIDVTIRNA
Lata_Latma-22    1170 YDKTALDIAIDAKF.SNIVEYLKTKSGKFRRES
Lata_Latma-23*   1279 ASDGITQAAALSIT.EKFEDVLNSLHNESAKEQ
Lata_Latma-24*?1315 EVHGKVYAALKSGRNSQIHQILCSSLNSISTLK
 Rnas_Human-1*     24 EDNHLLIKAVQNEDVDLVQQLLEGGANVNFQEE
 Rnas_Human-2*     58 GGWTPLHNAVQMSREDIVELLLRHGADPVLRKK
 Rnas_Human-3*     91 NGATLFILAAIAGSVKLLKLFLSKGADVNECDF
 Rnas_Human-4*    124 YGFTAFMEAAVYGKVKALKFLYKRGANVNLRRK
 Rnas_Human-5*    166 GGATALMDAAEKGHVEVLKILLDEMGADVNACD
 Rnas_Human-6*    202 GRNALIHALLSSDD4AITHLLLDHGADVNVRGE
a Rnas_Human-7*   238 RGKTPLILAVEKKHLGLVQRLLEQEHIEINDTD
 Rnas_Human-8*    272 DGKTALLLAVELKLKKIAELLCKRGASTDCGDL
 Rnas_Human-9*    300 DCGDLVMTARRNYDHSLVKVLLSHGAKEDFHPP
```

```
                                  aaaaaaaaa
                        t otLHhAh  tt thht LLt t t
  Lat2/Latma-1*    427 DIHRDLYNAAQVPY5SISRTLIQNGANVSETFE
  Lat2/Latma-2     464 LGRGAIHAAASAGNYDVGELLLNKDINLLEKAD
  Lat2/Latma-3     499 NGYTPLHIAADSNKNDFVMFLIGNNADVNVRTK
  Lat2/Latma-4     532 DLFTPLHLAARRDLTDVTQTLIDITEIDLNAQD
  Lat2/Latma-5     566 SGFTPLHLSISSTS.ETAAILIRNTNAVINIKS
  Lat2/Latma-6     600 VGLTPLHLATLQNNLSVSKLLAGKGAYLNDGDA
  Lat2/Latma-7     632 NGMTPLHYAAMTGNLEMVDFLLNQQYININAAT
  Lat2/Latma-8     668 KKWTPLHLAILFKKNDVAERLLSDENLNIRLET
  Lat2/Latma-9     701 GGINPLHLASATGNKQLVIELLAKNADVTRLTS
 Lat2/Latma-10     734 KGFSALHLGIIGKNEEIPFFLVEKGANVNDKTN
 Lat2/Latma-11     767 SGVTPLHFAAGLGKANIFRLLLSRGADIKAEDI
 Lat2/Latma-12     800 NSQMPIHEAVSNGHLEIVRILIEKDPSLMNVKN
 Lat2/Latma-13     834 RNEYPFYLAVEKRYKDIFDYFVSKDANVNEVDH
 Lat2/Latma-14     867 NGNTLLHLFSSTGELEVVQFLMQNGANFRLKNN
 Lat2/Latma-15     900 ERKTFFDLAIENGRLNIVAFAVEKNKVNLQAAH
 Lat2/Latma-16     933 RGKTILYHAICDSAKYDKIEIVKYFIEKLNESE
 Lat2/Latma-17     964 SECNPLHEAAAYAHLDLVKYFVQERGINPAEFN
 Lat2/Latma-18*    999 NQASPFCITIHGAPXEVVEYLSDKIPDINGKCD
 Lat2/Latma-19    1045 QENTPITVAIFANKVSILNYLVGIGADPNQQVD
 Lat2/Latma-20    1077 DGDPPLYIAARQGRFEIVRCLIEVHKVDINTRN
 Lat2/Latma-21    1111 ERFTALHAAARNDFMDVVKYLVRQGADVNAKGI
 Lat2/Latma-22    1144 DDLRPIDIAGEKAK2LQSSRFLRSGHSFQSNEI
 Lat2/Latma-23*   1260 RSAAEAQAEALIMT4NLLSGLIGDPIPDSIDFS
 Lat2/Latma-24*?1296 NVHSKIYKAIMSGRRSVISEMLCSFAEEYSKLN
  Li12_Caeel-1*   1021 ESPIKLHTEAAGSY.AITEPITRESVNIIDPRH
  Li12_Caeel-2    1052 HNRTVLHWIASNSSAEKSEDLIVHEAKECIAAG
  Li12_Caeel-3    1093 DENTPLMLAVLARRRRLVAYLMKAGADPTIYNK
  Li12_Caeel-4    1126 SERSALHQAAANRDFGMMVYMLNSTKLKGDIEE
  Li12_Caeel-5    1162 NGMTALMIVAHNEGXASAKLLVEKGAKVDYDGA
  Li12_Caeel-6    1206 KGRTALHYAAQVSNMPIVKYLVGEKGSNKDKQD
  Li12_Caeel-7    1240 DGKTPIMLAAQEGRIEVVMYLIQQGASVEAVDA
  Li12_Caeel-8*?1272 ATDHTARQLAQANNHHNIVDIFDRCRPEREYSM
  Mad3_Human-1*?     6 ERPQEWAMEGPRDG.LKKERLLDDRHDSGLDSM
  Mad3_Human-2      73 DGDSFLHLAIIHEEKALTMEVIRQVKGDLAFLN
  Mad3_Human-3     110 LQQTPLHLAVITNQPEIAEALLGAGCDPELRDF
  Mad3_Human-4     143 RGNTPLHLACEQGCLASVGVLTQSCTTPHLHSI
  Mad3_Human-5     182 NGHTCLHLASIHGYLGIVELLVSLGADVNAQEP
  Mad3_Human-6     216 NGRTALHLAVDLQNPDLVSLLLKCGADVNRVTY
  Mad3_Human-7     249 QGYSPYQLTWGRPS5QLGQLTLENLQMLPESED
  Notc_Drome-1    1901 CGLTPLMIAAVRGGXQVISDLLAQGAELNATMD
  Notc_Drome-2    1950 TGETSLHLAARFARADAAKRLFHAGADANCQDN
  Notc_Drome-3    1983 TGRTPLHAAVAADAMGVFQILLRNRATNLNARM
  Notc_Drome-4    2017 DGTTPLILAARLAIEGMVEDLITADADINAADN
  Notc_Drome-5    2050 SGKTALHWAAAVNNTEAVNILLMHHANRDAQDD
  Notc_Drome-6    2083 KDETPLFLAAREGSYEACKALLDNFANREITDH
  Ph81_Yeast-1*    379 SRKNVFHEAASCPE.KSRLFILDEALTTSKLSK
  Ph81_Yeast-2*    423 HSRVPLHYAAELGKLEFVHSLLITNLLEDVDPI
  Ph81_Yeast-3*    458 DSKTPLVLAITNNHIDVVRDLLTIGGANASPIE
  Ph81_Yeast-4*    506 VQFDPLNVACKFNNHDAAKLLLEIRSKQNADNA
  Ph81_Yeast-5*    556 TGLCTLHIVAKIGGDPQLIQLLIRYGADPNEID
  Ph81_Yeast-6*    591 NKWTPIFYAVRSGHSEVITELLKHNARLDIEDD
  Ph81_Yeast-7*    624 NGHSPLFYALWESHVDVLNALLQRPLNLPSAPL
  Phlb_Serli-1*?    41 DGHSNLALAQAVARGDTQGIHAQATQDRLRERG
  Phlb_Serli-2*     75 RQVTLLQWAVLSQQPDSVQALLDLGADPAAAGL
  Phlb_Serli-3*    108 DGNSALHTAAMLQDAQYLRLLLAEGAQMNVRNA
  Phlb_Serli-4*    142 TGATPLAAAVLAGREEQLRLLLAAGADTTLSDR
  Phlb_Serli-5*    175 LGDTPLHLAAKINRRTLALLLLQAGADARARNQ
  Phlb_Serli-6*    243 RLATQLRAAVNAHKKTAVQAVFLRQRGLFRVAG
  Swi4_Yeast-1     520 QGHTPLHWATAMANIPLIKMLITLNANALQCNK
  Swi4_Yeast-2*    550 CNKLGFNCITKSIF2NCYKENAFDEIISILKIC
  Swi4_Yeast-3*    590 NGRLPFHYLIELSV8PMIIKSYMDSIILSLGQQ
  Swi4_Yeast-4     641 IGNTPLHLSALNLNFEVYNRLVYLGASTDILNL
  Swi6_Yeast-1     317 HGNTPLHWLTSIANLELVKHLVKHGSNRLYGDN
  Swi6_Yeast-2*    350 MGESCLVKAVKSVN4GTFEALLDYLYPCLILED
  Swi6_Yeast-3*?   387 MNRTILHHIIITSG9YYLDILMGWIVKKQNRPI
  Swi6_Yeast-4     469 NGDTCLNIAARLGNISIVDALLDYGADPFIANK
     V1p_Rat-1*      1 .CDKEFMWALKNGDLDEVKDYVAKGEDVNRTLE
     V1p_Rat-2      33 GGRKPLHYAADCGQLEILEFLLLKGADINAPDK
     V1p_Rat-3      66 HHITPLSAVYEGHV.SCVKLLLSKGADKTVKGP
  Ycu1_Yeast-1*      1 .MNANIWVAASDGNLDRVEHILRESKGAMTPQS
  Ycu1_Yeast-2*     36 NGYTPMHAAAAYGHLDLLKKMCNEYNGDINVLD
  Ycu1_Yeast-3*     70 DGDTPLHHVEDVATARLIVEELGGDFTIRNVEG
  Ycu1_Yeast-4*    101 EGQTPYDSFVENGEDGELIEYMRIKSGVADVHG
  Ycu1_Yeast-5*?   126 SGVADVHGVDGVQGEGVIDSKLLEEFKDNVRYT
  Yjac_Ecoli-1*    274 INLPGLYLAINYGNADIVETIFNSLSETGYEGL
  Yjac_Ecoli-2*    322 NGFSGLFLAISRKDKNVVTSILNALPKLAATHH
  Yjac_Ecoli-3*    370 TSSHVLYHVMANGDADMLKIVLNALPLLIRTCH
  Yjac_Ecoli-4*    418 YGCPGLYLAMQNGHSDIVKVILEALPSLAQEIN
  Yjac_Ecoli-5*    466 ARDTGLFMAMQRGHMNVINTIFNALPTLFNTFK
  Yjac_Ecoli-6*    514 NEYPGLFSAIQHKQQNVVETVYLALSDHARLFG
  Yjac_Ecoli-7*    562 QKYSAFELAFEFGHRVIAELILNTLNKMAESFT
  Cact/Drome-1     228 DGDTPLHLACISGSVDVVAALIRMAPHPCLLNI
  Cact/Drome-2     265 VAQTPLHLAALTAQPNIMRILLLAGAEPTVRDR
  Cact/Drome-3     298 HGNTALHLSCIAGEKQCVRALTEKFGATEIHEA
  Cact/Drome-4     330 AHRQYGHRSNDKAVSSLSYACLPADLEIRNYDG
  Cact/Drome-5     361 RGERCVHLAAEAGHIDILRILVSHGADINAREG
  Cact/Drome-6     395 SGRTPLHIAIEGCNEDLANFLLDECEKLNLETA
  Cact/Drome-7     430 AGLTAYQFACIMNK.SRMQNILEKRGAETVTPP
  Clbp/Drome-1*     40 LEEKKFLLAVERGDMPNVRRILQKALRHQHINI
  Clbp/Drome-2*     78 LGRRALTLAIDNENLEMVELLVVMGVETKDALL
  Clbp/Drome-3*    105 TKDALLHAINAEFV.EAVELLLEHEELIYKEGE
  Clbp/Drome-4     152 PDITPLMLAAHKNNFEILRILLDRGAAVPVPHD
  Frk1/Drome-1*        NDVTPVYLAAQEGHLEVLKFLVLEAGGSLYVRA
  Frk1/Drome-2*        DGMAPIHAASQMGCLDCLKWMVSRWCTKCFEFG
  Frk1/Drome-3*        DGATPLHFAASRGHLSVVRWRLSRKLSLDKYGK
  Ph82/Yeast-1*        DGWTPFHIACSVGNLEVVKSLYDRPLKPDLNKI
  Ph82/Yeast-2*        QGVTCLHLAVGKKWFEVSQFLIENGASVRIKDK
  Ph82/Yeast-3*        SNQIPLHRAASVGSLKLIELLCGLGKSAVNWQD
  Ph82/Yeast-4*        NGWTPLFHALAEGHGDAAVLLVEKYGAEYDLVD
   Trp_Drome-1*     31 DVEKNFILSCERGDLPGVKKILEEYQGTDKFNI
   Trp_Drome-2*     69 MNRSALISAIENENFDLMVILLEHNIEVGDALL
   Trp_Drome-3*     96 VGDALLHAISEEYV.EAVEELLQWEETNHKEGQ
   Trp_Drome-4     143 VDITPLILAAHRNNYEILKILLDRGATLPMPHD
   G9a/Human-1*    440 FHPRQLYLSVKQGE2KVILMLLDNLDPNFQSDQ
   G9a/Human-2     475 SKRTPLHAAAQKGSVEICHVLLQAGANINAVDK
   G9a/Human-3     508 QQRTPLMEAVVNNHLEVARYMVQRGGCVYSKEE
   G9a/Human-4     541 DGSTCLHHAAKIGNLEMVSLLLSTGQVDVNAQD
   G9a/Human-5     575 GGWTPIIWAAEHKHIEVIRMLLTRGADVTLTDN
   G9a/Human-6     608 EENICLHWASFTGSAAIAEVLLNARCDLHAVNY
   G9a/Human-7     641 HGDTPLHIAARESYHDCVLLFLSRGANPELRNK
```

Fig. 4a.

```
pred. sec.struct.                    aaaaaaaaaa
consensus:            t oPLHhAh  tt thht LLt t
 Vc19_Vaccc-1      73 NVCHMYFTFFDVDT2HLFKLVIKHCDLNKRGNS
 Vc19_Vaccc-2     103 RGNSPLHCYTMNTR3SVLKILLHHGMRNFDSKD
 Vc19_Vaccc-3     137 DEKGHHYLIHSLSI2KIFDILTDTIDDFSKSSD
 Vc19_Vaccc-4     166 SKSSDLLLCYLRYK3SLNYYVLYKGSDPNCADE
 Vc19_Vaccc-5  ?  201 DELTSLHYYCKHISXRFIYAIIDYGANINAVTH
 Vc18_Vaccc-1       8 RFNNCGYHCYETILIDVFD.ILSKYMDDIDMID
 Vc18_Vaccc-2      41 ENKTLLYYAVDVNNIQFAKRLLEYGASVTTSRS
 Vc18_Vaccc-3      75 INTAIQKSSYQREN3RIVDLLLSYHPTLETMID
 Vc18_Vaccc-4     110 FNRDIRYLYPEPLF2IRYALILDDDFPSKVSMI
 Vc17_Vaccc-2      59 RGNNALHCYVSNKC4KIVRLLLSRGVERLCRNN
 Vc17_Vaccc-3      95 EGLTPLGAYSKHRY3QIVHLLISSYSNSSNELK
 Vc17_Vaccc-4     130 SNINDFDLSSDNIDLRLLKYLIVDKRIRPSKNT
 Vc17_Vaccc-5     169 LGLVDIYVTTPNPRPEVLLWLLKSECYSTGYVF
 Vc17_Vaccc-6     210 MCKNSLHYYISSHR7DVIKCLINNNVSIHGRDE
 Vc17_Vaccc-7     249 GGSLPIQYYWSFST3EIVKLLLIKDVDTCRVYD
 Vc17_Vaccc-8     284 VSPILEAYYLNKRF9EIVNLLIERRHTLVDVMR
 Vc17_Vaccc-9     331 SREYNHYIIDNILK7SIVQAMLINYLHYGDMRS
 Vc15_Vaccc-1      21 NMCHLYVKVCPSSL..LFRLFVECCDINKLVEG
 Vc15_Vaccc-2  ?   50 EGTTPLHCYLMNEGFESSVLKNLLKEYVMNTFN
 Vc09_Vaccc-2      36 DGETPLKAYVTKKN5DVVILLLSSVDYKNINDF
 Vc09_Vaccc-3      71 DFDIFEYLCSDNIDIDLLKLLISKGIEINSIKN
 Vc09_Vaccc-4     105 INIVEKYATTSNPNVDVFKLLLDKGIPTCSNIQ
 Vc09_Vaccc-6     175 MGKTVLYYYIITRS8DVINYLISHKKEMRYYTY
 Vc09_Vaccc-7     215 REHTTLYYYLDKCD3EIFDALFDSNYSGHELMN
 Vc09_Vaccc-8     242 YSGHELMNILSNYL9KIDNYIVDQLLFDRDTFY
Vc09_Vaccc-10     307 IQDLLLEYVSYHTV2NVIKCMIDEGATLYRFKH
Vc09_Vaccc-12     412 HGCSILYHCIKSHSVSLVEWLIDNGADINIITK
Vc09_Vaccc-13     445 YGFTCITICVILAD4EIAELYIKILEIILSKLP
 Vm01_Vaccc-1      17 NRNINFYTTMDNIM2EYYLSLYAKYNSKNLDVF
 Vm01_Vaccc-2      59 PSGNNYHILHAYCG6RFVEELLHRGYSPNETDD
 Vm01_Vaccc-3      97 DGNYPLHIASKINNNRIVAMLLTHGADPNACDK
 Vm01_Vaccc-4     130 HNKTPLYYLSGTDD3ERINLLVQYGAKINNSVD
 Vm01_Vaccc-5     166 EGCGPLLACTDPSE.RVFKKIMSIGFEARIVDK
 Vm01_Vaccc-6     198 FGKNHIHRHLMSDN3STISWMMKLGISPSKPDH
 Vm01_Vaccc-7     233 DGNTPLHIVCSKTV3DIIDLLLPSTDVNKQNKF
 Vm01_Vaccc-8     267 FGDSPLTLLIKTLS2HLINKLLSTSNVITDQTV
 Vm01_Vaccc-9     322 YDSTDFKMAVEVGSIRCVKYLLDNDIICEDAMY
Vm01_Vaccc-10  ?  356 SEYETMVDYLLFNH.FSVDSVVNGHTCMSECVR
Vm01_Vaccc-11  ?  405 PTSETMYLTMKAIE2KLDKSIIIPFIAYFVLMH
 Vhrp_Vaccc-1      29 HGHSALYYAIADNNVRLVCTLLNAGALKNLLEN
 Vhrp_Vaccc-2      60 ENEFPLHQAATLEDTKIVKILLFSGMDDSQFDD
 Vhrp_Vaccc-3      93 KGNTALYYAVDSGNMQTVKLFVKKNWRIFYGKT
 Vhrp_Vaccc-4     127 GWKTSFYHAVMLNDVSIVSYFLSEIPSTFDLAG
 Vhrp_Vaccc-5     160 ILLSCIHTTIKNGHVDMMILLLDYMTSTNTNNS
 Vhrp_Vaccc-6  ?  193 LFIPDIKLAIDNKDIEMLQALFKYDINIYSVNL
 Vb04_Vaccc-1     169 YGCTLLHRCIYHYK8ELIKILLNNGSDVDKKDT
 Vb04_Vaccc-2     209 YGNTPFILLCKHDINNVLFEICLENANIDSVDF
 Vb04_Vaccc-3     243 NRYTPLHYVSCRNKYDFVKLLISKGANVNARNK
 Vb04_Vaccc-4     276 FGTTPFYCGIIHGISLISKLYLESDTELEIDNE
 Vb04_Vaccc-5     305 DNEHIVRHLIIFDAVEVLDYLLSRGVIDINYRT
 Vb04_Vaccc-6     339 YNETSIYDAVSYNAYNLSVYLLNRNGDFETITT
 Vb04_Vaccc-7     372 SGCTCISEAVANNNKIIMEVLLSKRPSLKIMIQ
 Vb04_Vaccc-8     404 QSMIAITKAKQHNA.DLLKMCIKYTACMTDYDT
 Vb18_Vaccc-1  ?   22 NEIYTYFSHCNIDH3ELDFVVKNYDLNRRQHVT
 Vb18_Vaccc-2      56 TGYTALHCYLYNNY3DVLKILLNHDVNVTMKTS
 Vb18_Vaccc-3      91 SGRMPVYILLTRCC4DVVIDMIDKDKNHLSHRD
 Vb18_Vaccc-4  ?  125 RDYSNLLLEYIKSRXNIVSTLLDKGIDPNFKQD
 Vb18_Vaccc-5     166 DGYTALHYYYLCLAXRIISLFIQHGANLNALDN
 Vb18_Vaccc-6     217 CGNTPFHLYLSIEM4HMTKMLLTFNPNFKICNN
 Vb18_Vaccc-7     253 HGLTPILCYITSDY4ILVMLIHHYETNVGEMPI
b Vb18_Vaccc-8    187 PIDERRMIVFEFIK7DSITYLMNRFKNINIYTR
 Vb18_Vaccc-9     327 EGKTLLHVACEYNNTQVIDYLIRINGDINALTD
Vb18_Vaccc-10  ?  375 SPYTINCLLYILRY.IVDKNVIRSLVDQLPSLP
```

Fig. 4. Multiple alignment of selected ANK repeats. (**a**) Eukaryotes and prokaryotes; (**b**) vaccinia virus proteins. Repeats shown for the first time in an alignment of ANK-repeats are marked by a star. Repeats with a weak signal and with large deviations from the consensus (e.g., more than 10 mismatches[18]) are labeled by a question mark. Most of them have been included manually because of their location in between or next to other ANK repeats and thus have to be treated with caution at this point. The most divergent repeats occur in viruses[12] due to their faster mutation rate. The protein names were taken from SWISSPROT if available. The beginnings of the repeats in the respective proteins are given in the second column. Dots denote gaps and numbers in position 15 indicate inserts counted in amino acids (X indicates insertions between 10 and 13 animo acids). The central helix (a) as predicted by the PHD program is shown in the first line. A consensus line indicates conserved features (h, hydrophobic; t, turn-like or polar; o, S/T; capitals, conserved amino acids). The nomenclature of Goebel et al.[21] for vaccinia virus proteins is used except for VHRP, a host range protein (K1 in Fig. 3).

like ankyrins, and distinct transmembrane proteins such as the single membrane spanning *notch* or the plant potassium ion transport protein *aktl*[14] leave little room for the view of a motif with a highly specialized function. Nevertheless, despite their widespread occurrence and functional variety, most of the biochemically characterized proteins containing ANK repeats appear to be involved in protein–protein interactions. In addition to ankyrins, protein–protein interactions have been shown for another well-characterized but still rapidly emerging protein family (see, e.g., ref. 38): the transcription activators related to nuclear factor κB (NF-κB) and their inhibitors (IκB). Both the C-terminal ANK repeats of NF-κB-like precursors and those forming the IκB prevent transcription by binding to the activator domain (for reviews see refs. 39 and 40). The ANK repeats in a heteromeric purine-specific DNA binding protein (GABP) have been shown to mediate subunit contacts.[41] For other proteins such as black widow spider latrotoxin and latroinsectotoxin[36] as well as the heterotetrameric glutaminase,[37] protein–protein interactions via the ANK repeats are also very likely (Fig. 2).

ANK Repeats in Poxviruses

If the abundance of ANK repeats in functionally diverse eukaryotic and prokaryotic proteins is already noteworthy, their accumulation in poxviruses is even more surprising. The pattern searches showed that 13 out of the 198 "major" protein-coding regions of the complete vaccinia virus genome[21] contain ANK repeats. Homologous proteins have been found in a variety of other poxviruses such as shope fibroma, fowlpox, cowpox, and variola. Recently, the genomes of variola and related vaccinia viruses have been compared and the number of ANK repeats in variola virus might even be higher.[22] Interestingly, the ANK repeats are exclusively located near and within the inverted terminal repeats (Fig. 3). These regions contain mostly extracellular proteins which have probably been acquired from the hosts (Fig. 3); only two code for cytosolic protein kinases (data not shown). It might be a coincidence, but the location of receptors for interleukin-1 and tumor necrosis factor (TNF) next to the proteins containing ANK repeats (Fig. 3) is striking and suggests some functional relations. Both interleukin-1 and TNF are known cell stimuli which initiate NF-κB-directed transcription.[39,40] Concentration and recognition of these cell stimuli are one way to speed up transcription within infected cells. The viral ANK-repeats could support dissociation of cellular NF-κB/IκB by competition for IκB and could thus contribute to virus propagation within infected host cells.

DISCUSSION

Mutation Rates

In spite of the abundance of ANK repeats in current databases, only a few orthologous genes, mainly from higher animals, are available for evaluating mutation rates during evolution. It should be noted that ankyrin repeats are found in nearly all

```
          G TPLHhAht        thht Lht GA
175 ..LGDTPLHLAAKINRR..Twrccccrpgpmpgratsrasrssftsrkrrricrmtn*
175 ..aggypaasgged*pp..hLALLLLQAGADARARNQQGVAFQFY                 215
216   FSQTPAHLQNDELKAQFRELDKWLQGHRLATQLRAAVNAHKKTAVQAVFLRQRG..    313
      [           ANK-repeat                 ]
```

Fig. 5. A possible frameshift in the *Phlb* sequence. In *Phlb* the sequence similarity to ANK repeats abruptly drops within a ANK copy near the C-terminus and an amino acid composition unusual for proteins follows. A frameshift in position 191 would lead to a complete fifth repeat and also to a sixth although more divergent one. The proposed frame is not terminated within the sequenced region and suggests that the protein is longer than 313 amino acids.

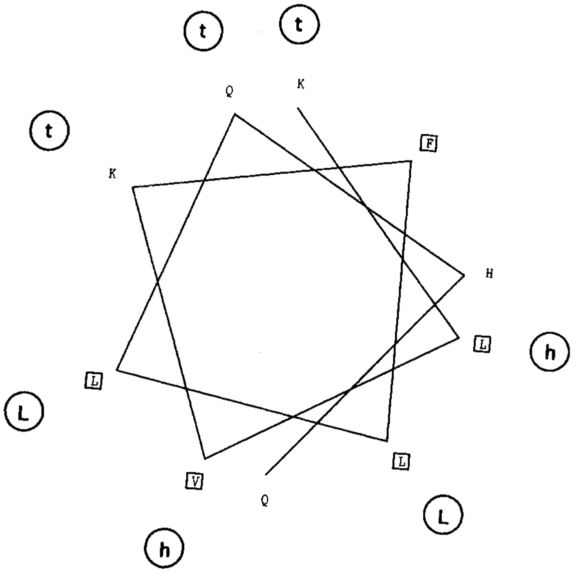


Fig. 6. Mapping of the conserved central positions onto a helical wheel. Hydrophobic positions are boxed. If a position is conserved, the corresponding symbol is printed next to the amino acid (circled; for nomenclature see the consensus line in Fig. 4). The conserved hydrophobic positions indicate a rather buried helix although three conserved polar residues are placed on one side in 3D. Due to the shortness of the repeat a tight packing of several ANK repeats is suggested.

TABLE I. Statistics of Variable External Repeats*

Proteins	N	N+1	C−1	C
1 Ankyrin	9	3	2	8
2 bcl3	13	2	1	7
3 cdc10	2	8	10	5
4 fem	9	3	2	5
5 Gabp	8	0	0	7
6 glp	7	4	3	13
7 Glsk	6	0	0	9
8 Kfb1	4	1	2	2
9 Latrot.	8	0	0	6
10 lin12	10	1	2	10
11 Mad3	5	1	1	10
12 notch/x	12	0	1	14
13 Ph81	4	2	2	13
14 Phlb	6	2	5	7
15 SWI4	1	7	9	4
16 Yjac	7	2	2	7

*Deviations from the final consensus pattern (Fig. 1) in several proteins are shown. N, N-terminal; N+1, second; C−1, penultimate; C, ultimate repeat. In most of the analyzed proteins the external repeats appear to be less conserved; notable exceptions are the transcription factors *SWI4, SWI6, cdc10,* and *res1* in which the ANK repeats apparently do not participate in protein–protein interactions but might be involved in DNA binding.[29]

phyla but that the surrounding regions of the respective proteins are very different. Is this phenomenon due to the incompleteness of the sequence data or is it a consequence of horizontal exchange of genetic material? In order to address this question, sequence similarities between available orthologous, but also between overall related paralogues (i.e., similar sequences that are apparently encoded by different genes in the respective organisms), have been compared (Table II).

Surprisingly, the ANK repeats show an extraordinarily high degree of similarity in all studied orthologues (Table II). It varies between 92 and 99% amino sequence identity for humans and rodents. These segments represent the most conserved parts of the respective proteins (Table II) and have very low amino acid exchange rates even when compared to a wide range of eukaryotic proteins.[2,42]

The amino acid similarity is much lower when comparing paralogues. Although in these comparisons it is impossible to measure mutation rates during evolution, the high conservation of ANK repeats relative to the other parts of the studied proteins is again obvious. These data also suggest strong functional and structural constraints in all cases and allow the evolution of the repeat to be traced back by phylogenetic studies despite the relative shortness of the repeat.

Evolution of ANK Repeats

The dendrogram shown in Figure 7 gives only a very rough estimate of the evolution of the repeats. Nevertheless, it mirrors several known features and also reveals some conclusions about the origin of particular proteins. For example, all proteins with an overall homology cluster together and corresponding repeats have the highest sequence similarity to each other. In case of the proteins related to NF-κB this means that an original ANK repeat has been duplicated several times before divergence into the different subfamilies. For *notch*-like proteins such a first duplication can be traced back to a point

TABLE II. Comparison of Pairwise Sequence Identities (in %) Between ANK Repeats (Upper Right) and Between Whole Proteins (Lower Left)*

Orthologues						Paralogues						
Erythrocyte ankyrin		1	2			Spider toxins		1	2			
Human	1	##	97			Latrotoxin	1	##	37			
Mouse	2	90	##			Latroinsectotoxin	2	36	##			
Bcl3		1	2			Transcription factors		1	2	3	4	5
Human	1	##	86			*Kfb1* human	1	##	38	48	36	48
Mouse	2	81	##			*Bcl3* human	2	na	##	36	37	41
						Mad human	3	na	na	##	40	38
pp50 (Kbfl)		1	2	3		*cactus* fruit fly	4	na	na	na	##	34
Human	1	##	93	75		*Lyt10* human	5	na	na	na	na	##
Mouse	2	88	##	74								
Chicken	3	73	71	##		Cell cycle proteins		1	2			
						cdc10 S. pombe	1	##	40			
Mad3		1	2	3		*SW16 S. sacchar.*	2	28	##			
Human	1	##	94	77								
Rat	2	92	##	77		Receptors		1	2			
Chicken	3	71	71	##		*Lin-12 C. elegans*	1	##	57			
						glp C. elegans	2	48	##			
notch		1	2	3	4							
Human	1	##	99	91	70	Photoinduction protein		1	2			
Rat	2	90	##	91	71	*trp* fruit fly	1	##	56			
Frog	3	74	74	##	70	*clpb* fruit fly	2	42	##			
Fruit fly	4	46	46	45	##							
						Human ankyrins		1	2			
						Erythrocytes	1	##	67			
						Brain	2	55	##			

*Mutation rates of orthologues and paralogues. Note that if the difference between the ANK region and the whole protein is rather small, the respective proteins are largely composed of ANK repeats. Otherwise, in all of the functional diverse proteins ANK repeats clearly represent the most conserved regions.

before divergence of invertebrates and vertebrates, because there are orthologues from vertebrates and *Drosophila* available. There is, however, no single case known where orthologues have been found among different phyla (animals, plants, fungi, protozoa, prokaryotes, archebacteria), e.g., 6 paralogues (containing ANK repeats) of yeast are known and 10 of humans, but there is not a single sequence with overall homology between the yeast and human proteins. Considering the very large number of proteins containing ANK repeats that have been sequenced so far, this is the first example where the evolution of so widespread a domain cannot be explained by gene duplication and exon shuffling. The first argument against this surprising observation is the current lack of data, e.g., a spectrin-based membrane skeleton appears to be present in plants,[43] but the constituent proteins have not yet been sequenced. Nevertheless, there are several facts that suggest irregularities including horizontal gene transfer.

Horizontal Gene Transfer

First, one of the prokaryotic proteins containing ANK repeats (YJAC in Fig. 2) has very similar, unusually long repeats and the dendrogram (Fig. 7) indicates a rather recent duplication of an unit originally 48 amino acids long. But where did this unit come from? Remarkably, YJAC belongs to a class of *E. coli* proteins that has been shown to be acquired by horizontal gene transfer because of their distinct codon usage[44] (I. Moszer and A. Danchin, personal communication). Another prokaryotic protein from *Chromatium vinosum*[11b] contains 9 repeats; the corresponding protein part (> 200 amino acids) is almost 40% identical to animal ankyrins. This similarity is comparable with that between paralogues of animal proteins (Table II), is unexpectedly high for a prokaryote/eukaryote comparison, and cannot be expected from the mouse/human ankyrin comparison[2,42] (Table II). Are these indications for another horizontal gene transfer? Whereas horizontal transfers occur frequently between prokaryotes, they represent a rather rare event between eukaryotes and prokaryotes.[7] Thus, the acquisition of a eukaryotic ancestor of YJAC by *E. coli* and the eukaryotic origin of the two other prokaryotic proteins still remains to be proven, but the ability of ANK repeats to spread among eukaryotic proteins horizontally can be shown in another case. The recently sequenced parts of a 88-kDa *Plasmodium* protein[45] are simply too similar to human erythrocyte ankyrin (98% amino acid identity) to be the result of

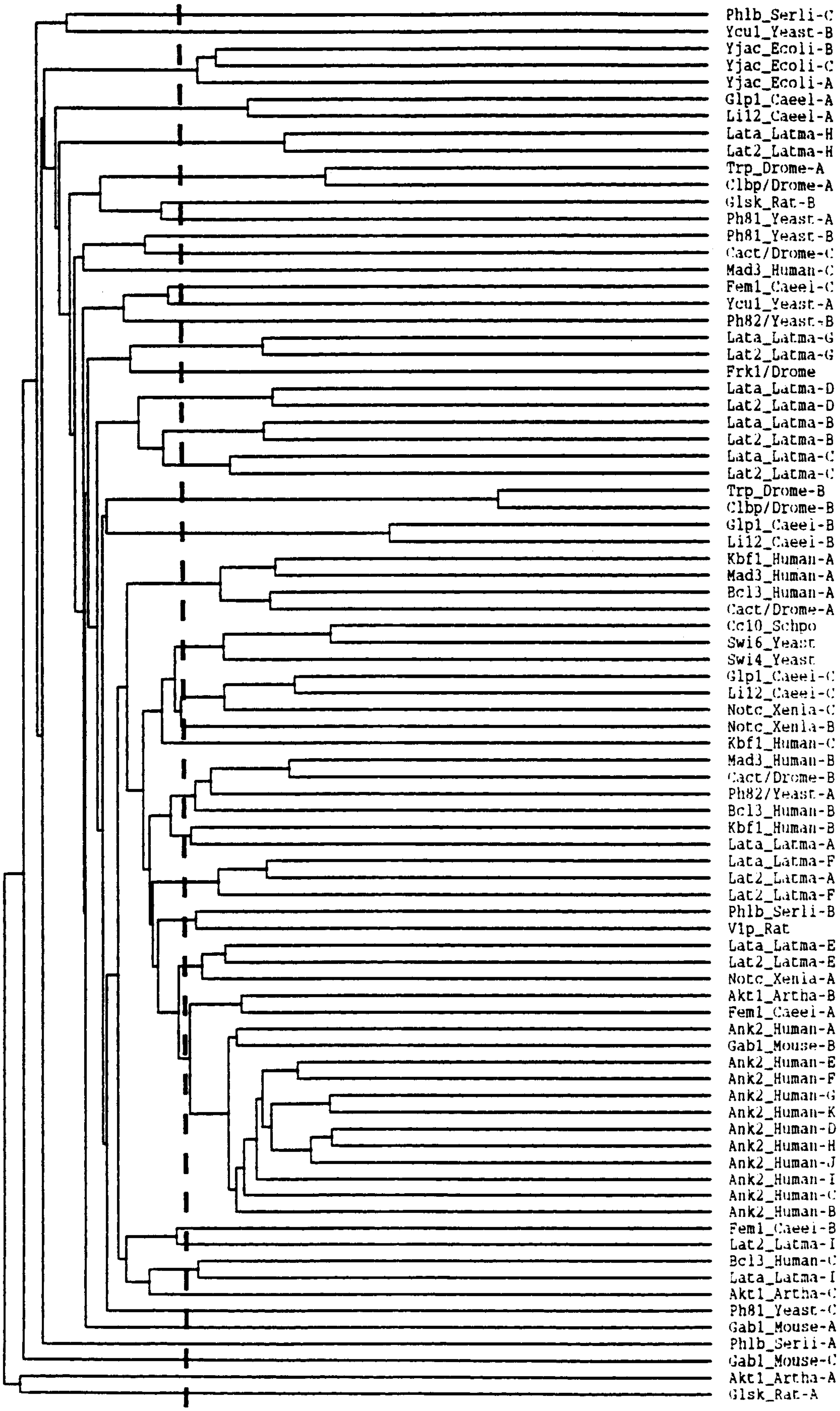

Phlb_Serli-C
Ycu1_Yeast-B
Yjac_Ecoli-B
Yjac_Ecoli-C
Yjac_Ecoli-A
Glp1_Caeel-A
Li12_Caeel-A
Lata_Latma-H
Lat2_Latma-H
Trp_Drome-A
Clbp/Drome-A
Glsk_Rat-B
Ph81_Yeast-A
Ph81_Yeast-B
Cact/Drome-C
Mad3_Human-C
Fem1_Caeei-C
Ycu1_Yeast-A
Ph82/Yeast-B
Lata_Latma-G
Lat2_Latma-G
Frk1/Drome
Lata_Latma-D
Lat2_Latma-D
Lata_Latma-B
Lat2_Latma-B
Lata_Latma-C
Lat2_Latma-C
Trp_Drome-B
Clbp/Drome-B
Glp1_Caeei-B
Li12_Caeei-B
Kbf1_Human-A
Mad3_Human-A
Bc13_Human-A
Cact/Drome-A
Cc10_Schpo
Swi6_Yeast
Swi4_Yeast
Glp1_Caeei-C
Li12_Caeei-C
Notc_Xenla-C
Notc_Xenla-B
Kbf1_Human-C
Mad3_Human-B
Cact/Drome-B
Ph82/Yeast-A
Bc13_Human-B
Kbf1_Human-B
Lata_Latma-A
Lata_Latma-F
Lat2_Latma-A
Lat2_Latma-F
Phlb_Serli-B
V1p_Rat
Lata_Latma-E
Lat2_Latma-E
Notc_Xenla-A
Akt1_Artha-B
Fem1_Caeei-A
Ank2_Human-A
Gab1_Mouse-B
Ank2_Human-E
Ank2_Human-F
Ank2_Human-G
Ank2_Human-K
Ank2_Human-D
Ank2_Human-H
Ank2_Human-J
Ank2_Human-I
Ank2_Human-C
Ank2_Human-B
Fem1_Caeei-B
Lat2_Latma-I
Bc13_Human-C
Lata_Latma-I
Akt1_Artha-C
Ph81_Yeast-C
Gab1_Mouse-A
Phlb_Serli-A
Gab1_Mouse-C
Akt1_Artha-A
Glsk_Rat-A

Fig. 7.

a long divergent evolution of both proteins.[45] Given the similarity between human and mouse ankyrin (97%, Table II), the different ankyrin variants in mammals, and assuming correct data, an acquisition of erythrocyte ankyrin by the protozoon *Plasmodium falciparum* is very likely.

Second, several genomes of different species allow first rough approximations about number and kind of proteins they contain. For example, more than 50% of all *E. coli* proteins are already stored in current databases (P. Rice and K. Rudd, personal communication) and two (out of 16) complete yeast chromosomes (III and XI) have already been sequenced, others will be finished soon.[46] In *E. coli,* only YJAC has been found yet to contain ANK repeats (A. Neuwald, personal communication). Considering the presence of at least 30 nonorthologous eukaryotic proteins (Fig. 2, Table I), numerous different proteins with ANK repeats would be expected to be present in *E. coli* if ANK repeats are indigenous.

In yeast, several proteins containing ANK repeats have been identified, but for none of these has a counterpart (orthologue) in animals been found yet. Green et al.[47] have proposed after extensive analysis of current sequence databases, that at least 60% of all genes either have evolved more recently than metazoan radiation or exhibit too fast a mutation rate to be detectable by homology searches. The latter cannot be true for ANK repeats because of their extremely slow mutation rate. If the proteins containing ANK repeats have evolved rather recently, what is the mechanism behind a so frequent insertion of ANK repeats into functionally diverse proteins? Exon shuffling[3] appears to be a valid argument only in animals and plants; the low number of introns in fungi and their absence in prokaryotes suggest another mechanism which remains to be identified. Horizontal gene transfer of certain domains among eukaryotes is a very rare event and cannot be the only reason for the presence of ANK repeats in such a variety of functionally diverse proteins.

CONCLUSION

In summary, sequence analysis revealed a surprising accumulation of ANK repeats in current databases. The presence of at least 639 repeats in 91 different proteins is comparable with the most widespread extracellular modules such as EGF-like domains (ca. 600 occurrences[48]) or fibronectin type III repeats (nearly 400 occurrences).[7] The comparison of all these ANK repeats allows a reliable characterization of structurally conserved features. Although a role in protein–protein interactions is suggestive, no explanation can be given so far for the spreading mechanism leading to both their widespread occurrence in functionally diverse proteins and to a remarkable abundance in poxviruses. However, since new members are continually being reported and it will soon be possible to compare complete genomes of different organisms, it should not be long before we can explore the role of horizontal gene transfers and domain shuffling in the creation of new proteins during evolution as suggested by this analysis.

NOTE ADDED IN PROOF

After acceptance of the manuscript the transcription factor *MBP1* related to *SW14/SW16* has been sequenced. It also contains 4 ANK repeats (C. Koch, T. Moll, M. Neuberg, H. Ahorn, K. Nasmyth. A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase. Science 261: 1551–1557, 1993). The presence of ANK repeats in *PHO81* has been noted independently by N. Ogawa et al. (Promoter of the *PHO81* gene encoding a 134 kDa protein bearing ankyrin repeats in the phosphatase regulon of *S. cerevisiae*. Mol. Gen. Genet. 238: 444–454, 1993.)

ACKNOWLEDGMENTS

The author is indebted to A. Neuwald, A. Danchin, and I. Moszer for information on YJAC, E. Koonin for hints regarding poxviruses, R.F. Doolittle for encouragement and helpful discussions, the referees for constructive comments, as well as C. Ouzounis, T. Gibson, and R. Wade for critical reading of the manuscript. This work was supported by grants from Deutsche Forschungsgemeinschaft and Boehringer Ingelheim Fonds.

REFERENCES

1. Bork, P. Mobile modules and motifs. Curr. Opin. Struct. Biol. 2:413–421, 1992.
2a. Doolittle, R.F. Reconstructing history with amino acid sequences. Prot. Sci. 1:191–200, 1992.
2b. Doolittle, R.F., Bork, P. Evolutionarily mobile modules in proteins. Sci. Amer. 269:50–56, 1993.
3. Rogers, J.H. The role of introns in evolution. FEBS Lett. 268:339–343, 1990.
4. Gilkes, N.R., Henrissat, B., Kilburn, D.G., Miller, R.C., Jr., Warren, R.A.J. Domains in microbial β 1,4-glycanases: Sequence conservation, function and enzyme families. Microb. Rev. 55:303–315, 1991.
5. Parkinson, J.S., Kofoid, E.C. Communication modules in bacterial signalling proteins. Annu. Rev. Genet. 26:71–112, 1992.
6. Saier, M.H., Reizer, J. Proposed uniform nomenclature for the proteins and protein domains of the bacterial phosphoenolpyruvate:sugar phosphotransferase system. J. Bacteriol. 174:1433–1438, 1992.

Fig. 7. Dendrogram of 88 selected eukaryotic and prokaryotic ANK repeats as produced by the PILEUP program of the GCG package.[17] For a better separation, two consecutive repeats have been fused, i.e., A means repeat 1 + 2, B, 3 + 4, etc. The vertical line indicates the first clearly wrong clustered pair of paralogues which can be taken as control for incorrect branching. Thus, all branching orders left of this line might be wrongly assigned. Nevertheless, several features including the striking internal similarities of repeats within ankyrin and also YJAC, however, appear to be significant. This might be due to recent duplication events or due to a slow molecular clock of these proteins.

7. Bork, P., Doolittle, R.F. Proposed acquisition of an animal protein domain by bacteria. Proc. Natl. Acad. Sci. U.S.A. 89:8990–8994, 1992.

8. Lux, S.E., John, K.M., Bennet, V. Analysis of cDNA for human erythrocyte ankyrin indicates a repeated structure with homology to tissue-differentiation and cell-cycle control proteins. Nature (London) 344:36–42, 1990.

9. Bennett, V. Ankyrins. J. Biol. Chem. 267:8703–8706, 1992.

10. Breeden, L., Nasmyth, K. Similarity between cell-cycle genes of budding yeast and fission yeast and the *Notch* gene of drosophila. Nature (London) 329:651–654, 1987.

11a. Michaely, P., Bennett, V. The ANK repeat: A ubiquitous motif involved in macromolecular recognition. Trends Cell Biol 2:127–129, 1992.

11b. Odata, M.M., Van Beeumen, J.J., Ambler, R.P., Meyer, T.E., Cusanovich, M.A. Nucleotide sequence of the heme subunit of flavocytochrome c from the purple phototropic bacterium, Chromatium vinosum. A 2.6 Kb DNA fragment contains two multiheme cytochromes, a flavoprotein, and a hemolog of human ankyrin. J. Biol. Chem. 268:14426–14431, 1993.

12. Massung, R.F., McFadden, G.M., Moyer, R.W. Nucleotide sequence analysis of a unique near-terminal region of the tumorigenic poxvirus, Shope fibroma virus. J. Gen. Virol. 73:2903–2911, 1992.

13. Robbins, J., Blondel, B.J., Callahan, D., Callahan, R. Mouse mammary tumor gene *int*-3: A member of the notch gene family transforms mammary epithel cells. J. Virol. 66:2594–2599, 1992.

14. Sentenac, H., Bonneaud, N., Minet, M., Lacroute, F., Salmon, J.M., Gaymard, F., Grignon, C. Cloning and expression in yeast of a plant potassium ion transport system. Science 256:663–665, 1992.

15. Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R., Sonnhammer, E. Comprehensive sequence analysis of the 182 predicted open reading frames of yeast chromosome III. Prot. Sci. 1:1677–1690, 1992.

16. Bairoch, A., Boeckmann, A. The SWISS-PROT protein sequence databank. Nucl. Acid Res. 20:2019–2022, 1992.

17. Devereux, J., Haeberli, P., Smithies, O. A comprehensive set of sequence analysis programs for the VAX. Nucl. Acid Res. 12:387–395, 1984.

18. Rohde, K., Bork, P. A fast, sensitive pattern-matching approach for protein sequences. CABIOS 9:183–189, 1993.

19. Gribskov, M., McLachlan, A.D., Eisenberg, D. Profile analysis: Detection of distantly related proteins. Proc. Natl. Acad. Sci. U.S.A. 84:4355–4358, 1987.

20. Pearson, W.R., Lipman, D. Improved tools for biological sequence comparisons. Proc. Natl. Acad. Sci. U.S.A. 85: 2444–2448, 1988.

21. Goebel, S.J., Johnson, G.P., Perkus, M.E., Davis, S.W., Winslow, J.P., Paoletti, E. The complete DNA sequence of vaccinia virus. Virology 179:247–266, 1990.

22. Shchelkunov, S.N., Blinov, V.M., Sandakhchiev, L.S. Ankyrin-like proteins of variola and vaccinia viruses. FEBS Lett. 319:163–165, 1993.

23. Coche, T. Prozzi, D., Legrain, M., Hilger, F., Vandenhaute, J. Nucleotide sequence of the PHO81 gene involved in the regulation of the repressible acid phosphatase gene in *Saccaromyces cerevisiae*. Nucl. Acid Res. 18:2176, 1990.

24. Zhou, A., Hassel, B.A., Silverman, R.H. Expression cloning of a 2-5A-dependent RNAase: A uniquely regulated mediator of interferon action. Cell 72:753–765, 1993.

25. Posfai, J., Roberts, R.J. Finding errors in DNA sequences. Proc. Natl. Acad. Sci. U.S.A. 89:4698–4702, 1992.

26. Givskov, M., Olsen, L., Molin, S. Cloning and expression in *E. coli* of the gene for extracellular phospholipase A1 from *Serratia liquefaciens*. J. Bacteriol. 170:5855–5862, 1988.

27. Phillips, A.M., Bull, A., Kelly, L.E. Identification of a *Drosophila* gene encoding a calmodulin-binding protein with homology to the *trp* phototransduction gene. Neuron 8:631–642, 1992.

28. Montell, C., Rubin, G.M. Molecular characterization of the *Drosophila trp* locus: A putative integral membrane protein required for phototransduction. Neuron 2:1313–1323, 1989.

29. Tanaka, K., Okazaki, K., Okazaki, N., Ueda, T., Sugiyama, A., Nojima, H., Okayama, H. A new cdc gene required for S phase entry of *Schizosaccharomyces pombe* encodes a protein similar to *cdc10* and *SWI4* products. EMBO J. 11:4923–4932, 1992.

30. Sidorova, J., Breeden, L. Analysis of the *SWI4/SWI6* protein complex, which directs G_1/S-specific transcription in *Saccharomyces cerevisiae*. Mol. Cell. Biol. 13:1069–1077, 1993.

31. Taoka, M., Yamakuni, T., Song, S.-Y., Yamakawa, Y., Seta, K., Okunuyama, T., Isobe, T. A rat cerebellar protein containing the *cdc10/SWI6* motif. Eur. J. Biochem. 207:615–620, 1992.

32. Smith, G.L., Chan, Y.S., Howard, S.T. Nucleotide sequence of 42 kbp of vaccinia virus strain WR from near the right inverted terminal repeat. J. Gen. Virol. 72: 1349–1376, 1991.

33. Rost, B., Sander, C. Structural prediction at better than 70% accuracy. J. Mol. Biol., in press, 1993.

34. Gibson, T. In EMBL research reports, HVA, Heidelberg, Germany, 1992.

35. Itoh, N., Phillips, S.E.V., Stevens, C., Ogel, Z.B., McPherson, M.J., Keen, J., Yadav, K.D.S., Knowles, P.F. Novel thioether bond revealed by a 1.7 Å crystal structure of galactose oxidase. Nature (London) 350:87–90, 1991.

36. Kiyatkin, N., Dulobova, I., Grishin, E. Cloning and structural analysis of α-latroinsectotoxin. Eur. J. Biochem. 213:121–127, 1993.

37. Shapiro, R.A., Farrell, L., Srinivasan, M., Curthoys, N.P. Isolation, characterization, and in vitro expression of a cDNA that endodes the kidney isoenzyme of the mitochondrial glutaminase. J. Biol. Chem. 266:18792–18796, 1991.

38. Iris, F.J.M. et al. Dense *Alu* clustering and a potential new member of the NF-κB family within a 90 kilobase HLA class III segment. Nature Genet. 3:137–145, 1993.

39. Nolan, G.P., Baltimore, D. The inhibitory ankyrin and activator *Rel* proteins. Curr. Opin. Gen. Dev. 2:211–220, 1992.

40. Blank, V., Kourilsky, P., Israel, A. NF-κB and related proteins: Rel/dorsal homologies meet ankyrin-like repeats. Trends Biochem. Sci. 17:135–140, 1992.

41. Thompson, C.C., Brown, T.A., McKnight, S.L. Convergence of Ets- and Notch-related structural motifs in a heteromeric DNA binding complex. Science 253:762–768, 1991.

42. Murphy, P.M. Molecular mimicry and the generation of host defense protein diversity. Cell 72:823–826, 1993.

43. Faraday, D.C., Spanswick, R.M. Evidence for a membrane skeleton in higher plants. FEBS Lett. 318:313–316, 1993.

44. Medigue C., Rouxel T., Vigier, P., Henaut, A., Danchin, A. Evidence for horizontal genetransfer in *E. coli* speciation. J. Mol. Biol. 222:851–856, 1991.

45. Suetterlin, B.W., Kappes, B., Jenoe, P., Franklin, R.M. An 88 kDa protein from *Plasmodium falciparum* is related to the band-3-binding domain of human erythrocyte ankyrin. Eur. J. Biochem. 207:457–461, 1992.

46. Goffeau, A., Nakai, K., Slonimski, P., Risler, J.-P. The membrane proteins encoded by yeast chromosome III. FEBS Lett. 325:112–117, 1993.

47. Green, P., Lipman, D., Hillier, L., Waterstone, R., States, D., Claverie, J.-M. Ancient conserved regions in new gene sequences and the protein database. Science 259:1711–1716, 1993.

48. Campbell, I.D., Bork, P. EGF-like modules. Curr. Opin. Struct. Biol. 2:385–392, 1993.

49. Xue, F., Cooley, L. *kelch* encodes a component of intercellular bridges in Drosophila egg chamber. Cell 72:681–693, 1993.