Evolutionary Relationships Among Proteins Probed by an Iterative Neighborhood Cluster Analysis (INCA). Alignment of Bacteriorhodopsins with the Yeast Sequence YRO2

Richard C. Graul[1] and Wolfgang Sadée[1,2]

Received June 30, 1997; accepted August 14, 1997

Purpose. Searching the existing databases for homologous sequences is essential to understanding a protein's structure and function. For a query sequence, its nearest neighbors can be identified by BLAST (basic local alignment search tool). However, a single query sequence is insufficient to define the entire neighborhood of related sequences, and multiple BLAST queries are needed. We describe here a program which permits automated and iterative BLAST analysis of an entire neighborhood of sequences and apply this to search for homologs of the bacteriorhodopsins outside the archaea phylum.

Methods. We have developed a Java program, 'Iterative Neighborhood Cluster Analysis' (INCA), which performs iterative BLAST searches, beginning with a single starter sequence, and proceeding with any other sequence achieving a predefined minimum alignment score. This results in a cluster of sequences where each sequence is related to at least one other sequence by the cutoff score, and additional lists of more distantly related sequences for each member of the cluster.

Results. Bacteriorhodospins had not been previously aligned with any other protein family with scores indicative of probable homology. Using INCA, we identified a probable homolog in yeast, YRO2_YEAST, also containing seven putative transmembrane domains. A finding of probable homology was supported by additional alignment strategies.

Conclusions. INCA is a useful tool to assess complete protein neighborhoods. With an increasing database, INCA can serve to detect the emergence of evolutionary links between even the most distantly related protein families. Identifying a homolog of the bacteriorhodopsins in yeast illustrates this approach but at the same time highlights the vast evolutionary distances between polytopic membrane proteins, such as the bacteriorhodopsins.

KEY WORDS: basic local alignment search tool; iterative neighborhood cluster analysis; bacteriorhodopsins; protein sequences.

INTRODUCTION

Sequence databases are rapidly expanding and currently include entire genomes for species in each of the major phyla, bacteria, archaea, and eukarya. This vast pool of information fundamentally changes research in all areas of the biosciences. In the pharmaceutical disciplines, drug discovery and development increasingly turn to genomics research to harvest information hidden within the multiplicity of gene families. Each protein subfamily, often containing several closely related members, is part of a larger family of homologous proteins, and further, a super-family consisting of distantly related genes that encode proteins with similar structural features. We now recognize that most drugs are likely to interact with multiple proteins in the body, even those previously considered specific, because of the presence of multiple genes related to the therapeutic target. Genomics research and bioinformatics address these issues by considering the universe of known sequences in an attempt to understand protein function and structure by considering protein evolution.

Establishing evolutionary relationships among distantly related proteins remains a major challenge. This is particularly difficult for polytopic membrane proteins containing multiple transmembrane domains (TMDs), such as G protein coupled receptors (GPCRs), ion channels, and transporters, representing the main drug targets in the body. Because they contain repetitive TMD segments, structural elements with restricted amino acid compositions, one cannot readily distinguish sequence similarities arising by convergence as opposed to alignments representing divergent sequences with common ancestry. Moreover, individual TMDs appear to represent separate and stable folding units within the membrane that may combine rather loosely into a tertiary polytopic protein (1); therefore, mutations could be more readily accommodated than is the case for soluble proteins, without loss of structural integrity, and the rate of divergence could be high. As a result, many gene families encoding polytopic membrane proteins share little sequence identity and seemingly stand alone in evolution despite extensive similarity in their secondary structure and membrane topology. For example, the bacteriorhodopsins, having a 7-TMD architecture resembling that of the GPCRs, have yet to be reliably aligned with any other protein family (2–4). Nevertheless, the bacteriorhodopsins have served as templates for molecular models of the GPCRs, although systematic differences between their structures have been noted (5–7). With a rapidly growing sequence database, it is likely that protein sequences are being discovered that are related to the bacteriorhodopsins and could thus provide a missing evolutionary link to known polytopic gene families, such as the GPCRs.

Searching the existing databases for protein homologs of a sequence of interest has become a routine first step in molecular biology research. Among the search algorithms, BLAST (basic local alignment search tool) serves to identify high scoring segment pairs (HSPs) among proteins that are similar in at least some portion of their sequence (8). As the number of available sequences rapidly expands, the number of chance alignments also increases. Consequently, deciding what BLAST scores are representative of homology and which proteins belong to a protein family becomes more difficult. Further, a single BLAST run has the limitation that it displays only the nearest neighbors of a single starter sequence. In order to have a complete view of the related sequences, one must run BLAST repeatedly with each of the neighbors, thus locating the neighbors of the neighbors (Fig. 1). Therefore, to define a protein family of nearest neighbors and establish more distant relationships requires multiple BLAST searches.

To facilitate this analysis, we have written a program, termed INCA (iterative neighborhood cluster analysis), which

[1] Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, University of California San Francisco, San Francisco, California 94143-0446.

[2] To whom correspondence should be addressed.

0724-8741/97/1100-1533$12.50/0

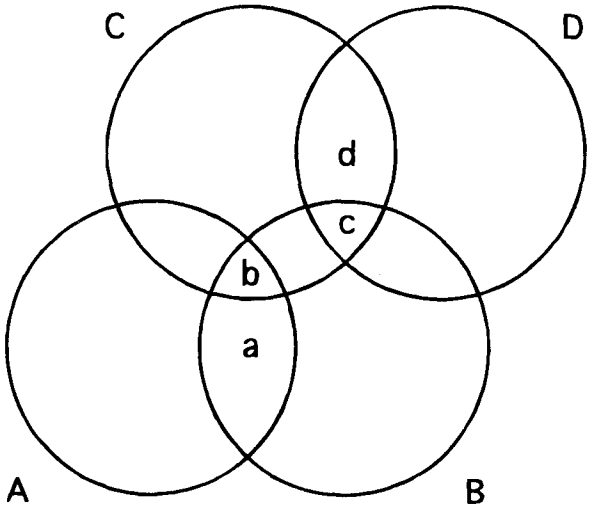

Fig. 1. Multiple query sequences yielding a cluster of neighborhoods. Neighborhoods obtained by BLAST (8) from single query sequences a–d are represented as circles A–D. There is no single best query sequence. Query sequence a finds neighborhoods A and B; b finds A, B, and C; c finds B, C, and D; d finds C and D.

permits automatic iterative access to the dynamically defined neighborhoods obtained from BLAST. For this program, we can specify stringency criteria to limit or expand the search for sequences with low similarity scores. Our program automatically searches, in turn, all sequences within a neighborhood. Further, INCA tabulates its results by defining a neighborhood cluster in which each sequence is related to at least one other sequence by a selected minimum similarity score, and further, provides lists of the more distant neighbors for each sequence in the cluster. This enables the rapid global assessment of sequence neighborhoods.

Our iterative search process, INCA, was applied to bacteriorhodopsins, a family of membrane proteins with seven transmembrane domains (TMDs) in *Archaebacteria* that serve as light driven H^+ or Cl^- pumps or as sensory proteins (5,9). Their molecular architecture is similar to that of the GPCRs, each containing seven TMDs, but sequences are too dissimilar to permit reliable alignments among them (2). Indeed, the bacteriorhodopsins have yet to be aligned with any other sequence with scores suggestive of probable homology, and therefore, represent an example of the lack of sequence similarity among polytopic membrane protein families despite similar topology. Thus, INCA could prove particularly useful for finding possible

```
                                                    10          20          30
BACT_NATPH MVGLTTLF-------------------------WLGAI---GMLVGTLAFAWAGRDAGS
           | . ..|.                         || ..   ..| |..  .   |.. .
YRO2_YEAST MSDYVELLKRGGNEAIKINPPTGADFHITSRGSDWLFTVFCVNLLFGVILVPLMFRKPVK
                     10         20         30         40         50         60

                   40         50         60                          70
BACT_NATPH GERRYYVTLVGISGIAAVAYVVMALGVGWVPVAE----------------RTVFAPRYID
           .   || | .. . . ..|| .|| ..||.|| .                 | .| .||..
YRO2_YEAST DRFVYY-TAIAPNLFMSIAYFTMASNLGWIPVRAKYNHVQTSTQKEHPGYRQIFYARYVG
                     70         80         90        100        110

                80         90        100        110        120        130
BACT_NATPH WILTTPLIVYFLGLLAGLDSREFGIVITLNTVVMLAGFAGAMVPGIERYALFGMGAVAFL
           |.|. |  .  ..||.|  ..... ......  .  . .| |... ... ...|  | .
YRO2_YEAST WFLAFPWPIIQMSLLGGTPLWQIAFNVGMTEIFTVCWLIAACVHSTYKWGYYTIGIGAAI
          120        130        140        150        160        170

              140        150        160        170                180
BACT_NATPH GLVYYLVGPMTESASQRSSGIKSLYVRLRNLTVILWAI-YPFIW-------LLGPPGVAL
            .   |...  . ...|.......... .......|| | ||  .       .| | ....
YRO2_YEAST VVCISLMTTTFNLVKARGKDVSNVFITFMSVIMFLWLIAYPTCFGITDGGNVLQPDSATI
          180        190        200        210        220        230

            190             200             210                    220
BACT_NATPH LTPTVDVAL-----IVYLDL-----VTKVGFGFIA-------------LDAAATLRAEHG
           .  ..|. .     .... |     ....|. | .            . ..|.......
YRO2_YEAST FYGIIDLLILSILPVLFMPLANYLGIERLGLIFDEEPAEHVGPVAEKKMPSPASFKSSDS
          240        250        260        270        280        290

                                                230
BACT_NATPH ES-----L-------------------------AGVDTDAPAVAD
           .|     |                          .. ..|... ..
YRO2_YEAST DSSIKEKLKLKKKHKKDKKKAKKAKKAKKAKKAQEEEEDVATDSE
          300        310        320        330        340
```

Fig. 2. Alignment of the entire primary structures of BACT_NATPH and YRO2_YFAST, using FASTA (18). A single dot denotes a conservative substitution, and a double dot identity. Dashed lines indicate gaps.

evolutionary links between the bacteriorhodopsins and other polytopic membrane proteins. For the bacteriorhodopsins we identify here a sequence with probable homology in yeast, YRO2 (10), and from there establish a possible link to a protein in the nematode *C.elegans*, YKQ3 (11), and further, yet another link to an expressed sequence tag (EST) in *Drosophila*. Additional links can be made to other known membrane proteins, but with alignments scores that are insufficient to establish a finding of probable homology at present. This paper describes our iterative search method, INCA, and the implications of our results linking bacteriorhodopsin of the archaea to sequences in eukaryotes.

EXPERIMENTAL APPROACH AND PROGRAM DESIGN

The Common Gateway Interface

The Common Gateway Interface (CGI) provides a mechanism to serve dynamically generated Web pages at a user's request. This is useful in applications, such as database searching, where the results are not known in advance. Typically, the user, with the aid of a Web browser, opens a Uniform Resource Locator (URL) on a server and is sent a form. The user fills in the form with the search parameters and submits it to the server. The server receives the search parameters via CGI. The server performs the requested database search and returns the results to the user as a Hypertext Markup Language (HTML) document. BLAST (8) represents an application at the National Library of Medicine/National Institutes of Health which uses CGI in this fashion. Access the URL http://www.ncbi.nlm.nih.gov/ for more information about this program. It is also possible to access CGI powered servers non-interactively using a computer program which simulates the user's request. Such a program is useful in applications which iteratively search a database. We have written a Java program which iteratively queries BLAST, searching for sequences which are similar to a user specified sequence.

ITERATIVE NEIGHBORHOOD CLUSTER ANALYSIS (INCA)

BLAST is a powerful tool for identifying protein neighbors (8). One provides BLAST with a query sequence, and BLAST returns a list of subject sequences which share primary structure similarities with the original query sequence. One can inspect the resultant list of subject sequences, with their accompanying scores and probabilities, and draw some cutoff indicative of significance. We call the resultant set of sufficiently similar subject sequences the nearest neighbors, or the neighborhood, of the query sequence.

The resultant neighborhood depends upon which specific query sequence is used to generate it. Suppose one generates a neighborhood **A** using sequence **a** as a query and finds subject sequence **b** to be a nearest neighbor (Fig. 1). Then one generates a second neighborhood **B** using sequence **b** as a query. The resultant two neighborhoods will overlap, but will not be identical. For example, subject sequence **c** may be sufficiently similar to query sequence **b** to be included in neighborhood **B**, but not sufficiently similar to query sequence **a** to be included in neighborhood **A**. If we only queried BLAST with sequence **a**, we would miss its relationship to sequence **c**; however, it may be possible to infer that since sequence **a** is related to sequence **b**, and sequence **b** is related to sequence **c**, then sequence **a** is related to sequence **c**. Therefore, if we are looking for distant evolutionary relationships, a single query may be insufficient.

A solution to this problem is to query BLAST multiple times, using an iterative neighborhood cluster analysis (INCA). The result of this analysis is a set of sequences composed from the union of one or more neighborhoods. We call this union the neighborhood cluster, or simply, the cluster. We have written a Java program, INCA, which automates this analysis. The program iteratively queries BLAST via CGI, and compiles the results. An Entrez query (http://www.ncbi.nlm.nih.gov/) (12) provides the primary structure of a sequence in FASTA format which we then use in a BLAST query to search for similar sequences. We use the BLAST probability result P(N) (13,14) as the criterion for a subject's inclusion into the cluster. A query-subject comparison must have a P(N) value less than or equal to a user provided cutoff ($\leq 10^{-6}$ to 10^{-15}) for the subject sequence to be included in the query neighborhood, and hence the cluster. To find relatively distant sequences that are still likely homologs, we use here a cutoff P(N) $\leq 10^{-6}$, that is the probability a given BLAST alignment, or the sum statistics (13,14) of several (N) BLAST alignments between the same sequence pair, could have occurred by chance in the database. During the first iteration, the cluster consists of all sequences below the cutoff value in alignments with the query sequence. During the second iteration, every sequence of the cluster is used in turn as a query, and any sequences scoring with P(N) $\leq 10^{-6}$ are added to the cluster, also to be used as a query. This procedure assures that the resultant set of sequences, each having at least one neighbor with P(N) $\leq 10^{-6}$, is independent of which starter sequence is used. As a result of the second iteration of this process, we complete the cluster and obtain a list of sequence neighbors for each member in the cluster, ordered by its P(N) value. To limit the lists of more distantly related sequences compiled for each cluster sequence, we choose a cutoff value of P(N) ≤ 0.1 or 0.01, or an arbitrary number of sequences, e.g., 100.

The results of the INCA search vary by selecting different BLAST search parameters. First, we select the alignment matrix, BLOSUM62 (15), being the default matrix of BLAST. To probe more distant relationships, the PAM120 or PAM250 matrices may be preferable for our application (8,15,16). Further, we can apply filters to mask query sequences of low complexity, i.e., acidic- or basic- or proline-rich regions of low information content. For example, the YRO2 sequence used here (10) as a query contains a highly charged C-terminus which is filtered out by SEG, the default filter for BLAST (17). Without any filtering, BLAST could identify numerous possibly unrelated nearest neighbors if sequences of low compositional complexity are present.

To limit the number of sequences in the cluster, we can increase the stringency of the INCA program (i.e., lower the cutoff). When we probe relationships for sequences that are members of large superfamilies containing numerous cloned genes (e.g., the GPCRs), the number of neighbors with low P(N) values may exceed 500, the arbitrary upper limit set by the BLAST program. In these cases, we can apply our INCA approach to search the predefined nearest neighbors in Entrez (12) using a modified program, INCA-Entrez. The Entrez neigh-

bors are also established by relying on a BLAST process with preset stringency criteria for each sequence in the database, but with no limits to the list of nearest neighbors (12). In this case, all nearest neighbors are already defined in the database, with no cutoff for the number of neighbors listed, and we simply proceed by using INCA for each predefined member in Entrez. However, we loose control over the stringency criteria that are critical for probing the most distantly related sequences. This may be compensated for by the large number of sequences probed, thereby, enhancing the chance to identify distant homologs. For the present application with bacteriorhodopsins, we did not use INCA-Entrez, but this program will also be available at the same Web site.

The INCA and INCA-Entrez programs will be freely available at the Web site http://sadee1.ucsf.edu/pub/software/java/inca/.

2D-Matrix Analysis, Hydropathy Analysis, and Alignments of Individual TMD Segments

The FASTA program of Pearson and Lipman (18) served to align two sequences with each other (see http://molbiol.soton.ac.uk/compute/align.html). To visualize the location of gaps or insertions, we also compared two sequences using a 2D-matrix analysis as described previously (19), using a flexible window (21–42 residues). We used the GES scale (20) to obtain hydropathy profiles, also as described (19). For separate analysis of each individual TMD segment (the 21-residue TMD, plus the adjoining loops or tails, maximally 25 residues), we determined the best alignment window between two TMD segments regardless of their location in the primary structure (19, and Graul, Babbitt, and Sadée, unpublished data). To account for amino acid bias in the TMDs (21–23) and to provide statistical evaluation for the pairwise alignments, alignments of TMD segments were analyzed by Monte Carlo calculations (1,000 randomizations of the second sequence of a pair, by applying a random number generator), yielding nSD_{MC} (number of standard deviations separating the BLOSUM62 score of the aligned two sequences from the mean of the BLOSUM62 scores obtained from all randomized sequence comparisons (19).

RESULTS AND DISCUSSION

We have written a Java program, termed INCA, which, given a starter sequence, automates multiple iterative BLAST analyses and filters the results to provide us with a resultant set of sequences. We then display an ordered listing of these sequences according to their distance (measured by the probability P(N) value) (13,14) from the starter sequence. A typical INCA run submits a hundred BLAST queries to define the entire neighborhood cluster of the protein sequence of interest. To illustrate this approach, we have applied INCA to the bacteriorhodopsins, 7-TMD membrane proteins of the archaea , which had yet to be aligned with any other protein family outside the archaea.

Starting with any query sequence of the bacteriorhodopsins, INCA yielded a cluster of 39 sequences in which each member aligns with at least one other member at $P(N) \leq 10^{-6}$. A nearly identical cluster results whether BLOSUM62 or PAM250 is used, which increases our confidence that the cluster represents a well defined nearest neighborhood. Any small differences between using BLOSUM62 and PAM250 related to smaller bacteriorhodopsin-like sequence fragments that did not affect any conclusions of our study. Searching for cluster sequences outside the archaea, we find four sequences belonging to the yeast genome. To verify that INCA would produce the same result even if a distant cluster member is used as the starter sequence, we selected the yeast protein YRO2 for the initial query, and indeed, the same cluster results. Shown in Table 1, a list of the neighborhood cluster displays the four yeast sequences and selected bacteriorhodopsin sequences (to avoid redundancies among very closely related proteins). These results suggest that the four yeast sequences may be related to the bacteriorhodopsins, with P(N) values just below the cutoff ($P(N) < 10^{-6}$). Among the yeast sequences are two nearly identical heat shock proteins (e.g., HSP30) which are highly expressed in yeast under stress conditions, including heat. Each of the four proposed proteins shares a 7-TMD topology with the bacteriorhodopsins, as predicted from hydropathy analysis.

To illustrate the differences between INCA and a single BLAST, we also show the results of the single BLAST analysis with YRO2 as the query in Table 2A. BLAST identified only 10 sequences with $P(N) < 0.01$, in contrast to the much larger cluster and extended neighborhood found through INCA. Whereas YRO2 does find the bacteriorhodospins it is most closely related to, most bacteriorhodopsins fail to show up in this list.

Inclusion of the yeast sequences with the bacteriorhodopsin cluster does not prove ancestral relationships. To investigate this further, we focus on a detailed analysis of the best alignment between yeast sequences and bacteriorhodopsins, i.e., YRO2_YEAST versus BACT_NATPH, with a sum P(2) $4.9 \cdot 10^{-7}$. In this case, BLAST identifies three local alignments, shown in Table 2, two of which are used for calculating the P(2) value. These three HSPs occur in consecutive order in the primary sequence, an important additional criterion in evaluating possible homology which is not reflected in the P(N) value. To illustrate the location of these alignments relative to the predicted TMDs, we show a dot plot (19) comparing YRO2_YEAST with BACT_NATPH in Fig. 3. The two darkest lines (indicating strongest sequence similarity) are consecutive with only one gap between them. The calculated hydropathy plots for the two sequences, given in Fig. 4, suggest a similar 7-TMD profile, but more importantly, reveal that at least four of the seven TMD segments can be aligned with BLOSUM62 scores above 20, in the same order in which they appear in both sequences. This result indicates that the order of TMDs has been conserved in the process of evolution, if these proteins are indeed homologs.

Evaluating homology among polytopic membrane proteins, however, presents unique difficulties because of the restricted amino acid distribution in the TMDs and their repetitive topology (21–23). We have developed an approach that relies on the analysis of each TMD segment (TMD plus adjoining loops) separately (19, and Graul, Babbitt, and Sadée, to be published). By finding the best alignments between each TMD segment, regardless of its location in the primary structure, we can establish statistical limits for unexpected similarities. Using BLOSUM62, TMD segments considered to be unrelated, i.e., those of G protein coupled receptors and transporters, yield scores of 11.6 ± 6.6 (21 residues minimum length per sequence

Table 1. Iterative Neighborhood Cluster Analysis (INCA) with YRO2 as the Starter Sequence

A. We used BLOSUM62 and the following databases: Non-redundant GenBank CDS translations + PDB + SwissProt + SPupdate + PIR; 256, 092 sequences. The SEG filter was applied to suppress alignments in the polar tail of YRO2 with an amino acid composition of low complexity (17). INCA identified 340 total neighbors (P(N) <1) of which 39 were inside the neighborhood cluster defined with a probability P(N) of $<10^{-6}$. To assess overall relationships within the cluster, we include two probability values, the first comparing a reference sequence of the cluster (identified by the locator number) to the query sequence, and the second, comparing the newly added cluster member to the reference sequence. The second P(N) value provides the best alignment for that member within the cluster. Only selected bacteriorhodopsins are.

```
displaying cluster with P(N) >10e-6 (edited to remove redundancies)
00586913    (7.3e-218)    <>    00586913    (7.3e-218)    YRO2 PROTEIN
00586913    (7.3e-218)    ->    01122353    (1.9e-159)    (Z68196) unknown [Saccharomyces]
00586913    (7.3e-218)    ->    00485481    (2.2e-38 )    heat shock protein HSP30-yeast
00586913    (7.3e-218)    ->    00140468    (1.0e-38 )    30 KD HEAT SHOCK PROTEIN
00586913    (7.3e-218)    ->    01168615    (4.9e-07 )    SENSORY RHODOPSIN II (SR-II2
01168615    (4.9e-07 )    ->    01363465    (6.6e-15 )    pSR-II protein-Natronobacterium p.
01168615    (4.9e-07 )    ->    00235918    (3.0e-33 )    (S56354) archaerhodopsin-2 = retinal
01168615    (4.9e-07 )    ->    01085725    (1.7e-51 )    cruxrhodopsin-Haloarcula sp.
01168615    (4.9e-07 )    ->    01085724    (3.8e-12 )    cruxhalorhodopsin-Haloarcula sp.
01168615    (4.9e-07 )    ->    01527138    (3.2e-67 )    (U62676) sensory rhodopsin II [Halob.]
01168615    (4.9e-07 )    ->    00461613    (2.0e-30 )    SENSORY RHODOPSIN I (SR-I)
01168615    (4.9e-07 )    ->    00461612    (5.3e-21 )    BACTERIORHODOPSIN (BR)
01168615    (4.9e-07 )    ->    00461608    (4.9e-10 )    HALORHODOPSIN (HR)
01168615    (4.9e-07 )    ->    00231626    (7.9e-34 )    ARCHAERHODOPSIN 2 PRECURSOR
01168615    (4.9e-07 )    ->    00114812    (3.6e-23 )    SENSORY RHODOPSIN I
01168615    (4.9e-07 )    ->    00114807    (4.2e-30 )    ARCHAERHODOPSIN 1 PRECURSOR
01168615    (4.9e-07 )    ->    01217898    (5.9e-31 )    (D83748) csR3 [Haloarcula vallismortis]

displaying nearest neighbors with P(N) >10e-6(BLOSUM62)
 1.          00586913          (7.3e-218)          YRO2 PROTEIN
11.          00465754          (0.0097  )          YKQ3_CAEEL, HYPOTHETICAL 42.1 KD PROTEIN
```

B. To identify more distant relationships, INCA was again run with YRO2 and PAM250. This resulted in a cluster 37 sequences with P(N) $<10^{-6}$, and 2157 seqeunces with P(N) $>10^{-6} < 1$.

```
displaying nearest neighbors with P(N) >10e-6(BLOSUM62)
 1.          00586913          (5.7e-147)          YRO2 PROTEIN
 7.          01771316          (0.0070  )          (Y08870) serotonin transporter
13.          00400630          (0.016   )          SODIUM-DEPENDENT SEROTONIN TRANSPORTer
15.          01584496          (0.042   )          chemosensory receptor [Caenorhabditis]
16.          00113056          (0.043   )          NEURONAL ACETYLCHOLINE RECEPTOR PROTEIN

 2.          01122353          (5.8e-112)          (Z68196) unknown [Saccharomyces]
11.          01352545          (0.0098  )          NADH-UBIQUINONE OXIDOREDUCTASE CHAIN
12.          00132206          (0.013   )          G PROTEIN-COUPLED RECEPTOR RDC1

 3.          00140468          (2.3e-33 )          30 KD HEAT SHOCK PROTEIN
 5.          01084988          (0.00070 )          G5 (ND3) protein-Sauroleishmania
 7.          00576765          (0.058   )          (U15681) cytochrome b [Myrmecia]

 5.          01168615          (1.2e-09 )          SENSORY RHODOPSIN II (SR-II)
37.          01084141          (0.0090  )          Na+/H+ antiporter NhaA-Vibrio p.

17.          00461608          (0.049   )          HALORHODOPSIN (HR)
33.          01718473          (0.013   )          (U78676) cytochrome oxidase subunit
25.          00461610          (0.10    )          BACTERIORHODOPSIN (BR)
35.          01816522          (0.011   )          (U74650) CysZ [Escherichia coli]
```

pair (n = 25,473). Similarly, aligning the seven TMDs of YRO2 with each of the TMDs contained in 100 transporter proteins (each with 9–14 TMDs) gave BLOSUM62 alignment scores of 11.3 ± 7.0 (n = 8,491, highest score = 45) (Fig. 4). Thus sequence pairs scoring above 51–53 fall outside 6 × S.D. of the mean and are indicative of probable homology. Using these benchmark values, we compared the TMD segments of YRO2 with those of BACT_NATPH. The best alignments, shown in Table 3, scored with BLOSUM62 values of 51.0 and 54.0 for TMD2/TMD2 and TMD3/TMD3, respectively. Scrambling one sequence of each pair gave Monte Carlo S.D. values (nSD_{MC}) of 7.6 and 6.5 (Table 3), in the range of probable homology. (The nSDMC value serves to accounts for possible amino acid bias in TMDs; a low value would result if only a few lipophilic amino acids occur repeatedly in the TMD sequence.) Finally, the greatest sequence similarities occurred in regions that are highly conserved among the bacteriorhodopsins (TMDs 2, 3, and 6) (9), suggesting that these domains serve some important function that may have been conserved in the yeast protein. On the other hand, single residues characteristic of the func-

Table 2. Single BLAST Analysis Using YRO2_YEAST PROTEIN as the Query

A. We used BLOSUM62 and the same databases as in Table 1. The BLAST identified 10 sequences with P(N) <0, 01 (shown below). Compare to INCA results in Table 1.

Sequences producing High-scoring Segment Pairs:		High Score	Smallest Sum Probability P(N)	N
gi|586913|sp|P38079|YRO2_YEAST	YRO2 PROTEIN /gi|626980	1566	7.4e-218	1
gi|1122353|gnl|PID|e213800	(Z68196) unknown [Sacchar.]	1154	1.9e-159	1
gi|140468|sp|P25619|HS30_YEAST	30 KD HEAT SHOCK PROTEIN	118	1.0e-38	4
gi|485481|pir||S31848	heat shock protein HSP30	118	2.3e-38	4
gi|1168615|sp|P42196|BACT_NATPH	SENSORY RHODOPSIN II	83	4.9e-07	2
gi|1877020	(D50848) archaerhodopsin	74	5.8e-07	2
gi|114807|sp|P19585|BAC1_HALS1	ARCHAERHODOPSIN 1 PRECURSOR	74	2.2e-06	2
gi|461610|sp|P33969|BACR_HALHM	BACTERIORHODOPSIN (BR)	78	0.00026	2
gi|231626|sp|P29563|BAC2_HALS2	ARCHAERHODOPSIN 2 PRECURSOR	67	0.0029	2
gi|235918	(S56354) archaerhodopsin	64	0.0075	2
gi|465754|sp|P34298|YKQ3_CAEEL	HYPOTHETICAL 42.1 KD PROTEIN ...	64	0.0098	2

B. High-scoring sequence pair (HSP) BLAST alignments of YRO2 (query) with BACT_NATPH (subject)

```
 Score = 83 (39.1 bits), Expect = 4.9e-07, Sum P(2) = 4.9e-07
 Identities = 19/108 (17%), Positives = 43/108 (39%)

Query:       110        RQIFYARYVGWFLAFPWPIIQMSLLGGTPLWQIAFNVGMTEIFTVCWLIAACVHSTYKWG 169
                        R +F  RY+ W L   P   +  + LL G     +     + +  +  +      A V     ++
Sbjct:        66        RTVFAPRYIDWILTTPLIVYFLGLLAGLDSREFGIVITLNTVVMLAGFAGAMVPGIERYA 125

Query:       170        YYTIGIGAAIVVCISLMTTTFNLVKARGKDVSNVFITFMSVIMFLWLI 217
                         + +G  A + +   L+         R   + ++++  +++ LW I
Sbjct:        126        LFGMGAVAFLGLVYYLVGPMTESASQRSSGIKSLYVRLRNLTVILWAI 173

 Score = 62 (29.2 bits), Expect = 4.9e-07, Sum P(2) = 4.9e-07
 Identities = 11/31 (35%), Positives = 18/31 (58%)

Query:        61        DRFVYYTAIAPNLFMSIAYFTMASNLGWIPV 91
                        +R  Y T +  +   ++AY MA +GW+PV
Sbjct:        33        ERRYYVTLVGISGIAAVAYVVMALGVGWVPV 63

 Score = 37 (17.4 bits), Expect = 0.0018, Sum P(2) = 0.0018
 Identities = 11/46 (23%), Positives = 20/46 (43%)

Query:        26        FHITSRGSDWLFTVFCVNLLFGVILVPLMFRKPVKDRFVYYTAIAP 71
                        +++T G   +  V  V + GV  VP+  R     R+++    P
Sbjct:        36        YYVTLVGISGIAAVAYVVMALGVGWVPVAERTVFAPRYIDWILTTP 81
```

C. HSPs of YRO2 (query) with YKQ3_CAEEL (subject)

```
 Score = 64 (30.1 bits), Expect = 0.0099, Sum P(2) = 0.0098
 Identities = 14/44 (31%), Positives = 26/44 (59%)

Query:        27        HITSRGSDWLFTVFCVNLLFGVILVPLMFRKPVKDRFVYYTAIA 70
                        H +SRG+  +F+VF+ L+  ++++P+ RK V     Y  +A
Sbjct:         5        HASSRGNISIFSVFLIPLIAYILILPGVRRKRVVTTVTYVLMLA 48

 Score = 53 (24.9 bits), Expect = 0.0099, Sum P(2) = 0.0098
 Identities = 16/47 (34%), Positives = 22/47 (46%)

Query:       240        FYGIIDLLILSILPVLFMPLANYLGIERLGLIFDEEPAEHVGPVAEK 286
                        FY I   + IL +L  L + +  +  I  L    D    EHVGP  +K
Sbjct:       248        FYLIFAIGILCVLCGLGLGICEHWRIYTLSTFLDASLDEHVGPKWKK 294
```

tional sub-families among the bacteriorhodopsins (9) differ in YRO2. Taken together, these results strongly support a finding of probable homology between the bacteriorhodopsins and the yeast protein YRO2.

Three closely related yeast homologs of YRO2 are included with the bacteriorhodopsin cluster, e.g., a 30 kD heat shock protein, each with a deduced 7-TMD topology (24). The heat shock protein is abundantly expressed in yeast under heat stress, but its function remains unknown. Possibly, it assists in protein translocation across membranes, but at present, it is not possible to draw any structure-function relationships from our alignments. However, having found putative yeast homologs of the bacteriorhodopsins, we can probe further into their sequence neighborhoods in order to find additional links to protein families with known functions. Because the C-terminus of YRO2 is highly charged, and therefore, contains repetitive elements

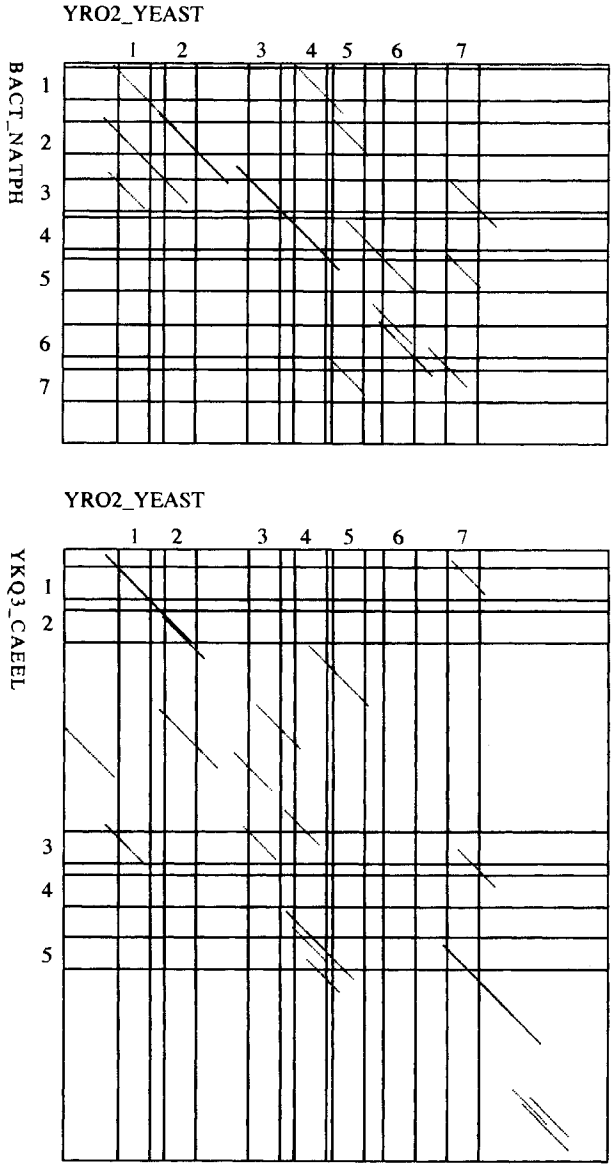

Fig. 3. 2D-Matrix dot plots comparing BACT_NATPH, YRO2_YEAST, and YKQ3_CAEEL. Diagonal lines indicate similarity exceeding a BLOSUM62 score of 20 (21–42 moving window) (19). Higher scoring regions are indicated by darker diagonal lines. The numbers on the x and y axes represent the TMD in consecutive order (deduced from hydropathy analysis) (20). The vertical and horizontal lines indicate the TMD boundaries.

consisting of charged amino acids, a filter is applied (SEG) (17) to exclude multiple hits with other sequences sharing this feature with YRO2, which is absent in the bacteriorhodopsins. Using BLOSUM62, the nearest neighbor outside the cluster is a hypothetical nematode protein, YKQ3_CAEEL, scoring with P(N) 0.0097 against YRO2 (Table 2A). Interestingly, the calculated hydropathy profile indicates only five TMDs, with the strongest similarity between YRO2 and YKQ3 occurring in their respective first and last TMDs (Figs. 3 and 4, BLAST alignment shown in Table 2C). Therefore, if these proteins are related, a 2-TMD segment has been inserted or deleted in the course of evolution between them. Analyzing each individual TMD segment separately gave two alignments with high scores, shown in Table 3. These result support a finding of probable homology for both TMD alignments separately, which is strengthened by occurrence in sequential order within the same sequences. However, the overall alignments between YRO2 and YKQ3 failed to reach our preselected cutoff for P(N) of 10^{-6}, and therefore, this alignment supports a finding of homology less strongly than that between YRO2_YEAST and BACT_NATPH.

To explore further the sequence neighborhood of the bacteriorhodopsin cluster (including YRO2 and related yeast sequences), we ran INCA with YRO2 as the starter sequence, employing SEG as the filter, and using PAM250 (the matrix most suitable for distant relationships). Whereas the bacteriorhodopsin-YRO2 cluster (defined by a P(N) of 10^{-6}) was largely unaffected, additional sequences were identified as nearest neighbors of the cluster, with P(N) values <0.1 or 0.01. An edited list of these neighbors is also shown in Table 1B. Interestingly, a variety of polytopic membrane proteins scored with intermediate P(N) values, regardless of whether they represent 7-TMD structures or proteins with a different number of TMD segments. Even though the bacteriorhodopsins share structural and functional features with the GPCR, only one putative GPCR attained sufficient scores to appear in this list, the putative G protein coupled receptor RCD1. Other proteins identified include serotonin transporters, a Na^+/H^+-antiporter, cytochrome oxidase subunits, and subunits of nicotinic ion channels. The BLAST scores, and those between individual TMD segments, reveal unexpected similarities, but they fall short of supporting a finding of probable homology.

The non-redundant databases searched by INCA (see Table 1A) do not contain all currently known sequences. In particular, databases consisting of expressed sequence tags (EST; non-redundant Database of GenBank EST Division), i.e., partial sequences of cDNA extracted from various tissues and species, exceed 1×10^6 entries. Whereas we did not find any additional related sequences with bacteriorhodopsins and YR02 as the query, INCA revealed a sequence tag highly similar to YKQ3, isolated from *Drosophila* (P(3) 2.1×10^{-22}; gi|2153031|gb|AA441153|AA441153 LD16050.5prime LD *Drosophila melanogaster* cDNA clone LD16050 5′, Length = 741). The probability score (P3; representing 3 HSPs) supports a finding of homology; therefore, by using the EST database we have extended an evolutionary link from *C. elegens* to *Drosophila*. While it is premature at this point to pursue this link further, it illustrates the use of INCA and the growing sequence databases to search for evolutionary links. However, considering the large database already available and the paucity of possible links found, these results illustrate the vast evolutionary distances among polytopic membrane proteins, in particular between bacteriorhodopsins and GPCRs.

Our results suggest two possible hypotheses to account for these distant relationships. First, membrane proteins identified here with intermediate scores are of different ancestry, and the similarities observed among them are the result of convergent evolution, to accommodate the structural requirements for repeat TMD segments. Second, polytopic membrane proteins may be related to each other and have evolved from fragments containing one or more TMDs. Because each TMD appears to represent a thermodynamically stable folding unit

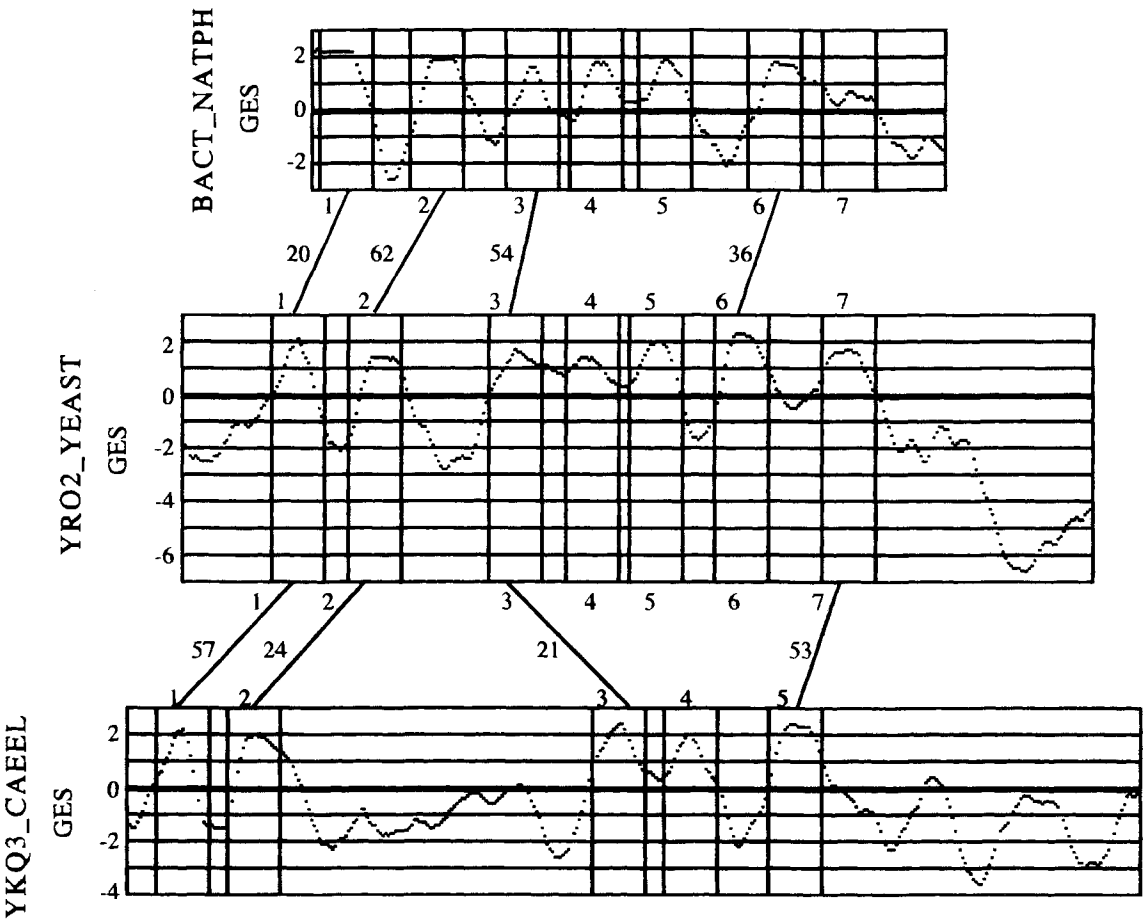

Fig. 4. Hydropathy profiles of BACT_NATPH, YRO2_YEAST, and YKQ3_CAEEL. We used the GES scale for hydropathy analysis (20). The vertical lines represent the TMD boundaries, and the TMDs are numbered consecutively as in Fig. 3. Lines connecting the hydropathy profiles represent alignments of individual TMD segments that gave BLOSUM62 scores >20. The actual score is also provided for each of these TMD alignments.

Table 3. Alignments of Individual TMD Segments

The primary sequences were divided into TMD segments (the hydropathy defined 21-residue TMDs, plus adjoining loops or tails containing maximally 25 residues), and each segment was compared against each other segment, regardless of location in the primary structure. To identify the highest scoring alignment pairs for each segment, we used BLOSUM62. nSDMC represent the number of standard deviations obtained from a Monte Carlo calculation where the second sequence was scrambled 1,000 times.

BACT_NATPH versus YRO2_YEAST

```
BACT_NATPH tmd02 = ERRYYVTLVGISGIAAVAYVVMALGVGWVPV
YRO2_YEAST tmd02 = DRFVYYTAIAPNLFMSIAYFTMASNLGWIPV
length = 31, ident.= 35%, score = 62.0, nSDMC = 7.6
BACT_NATPH tmd03 = RTVFAPRYIDWILTTPLIVYFLGLLAG
YRO2_YEAST tmd03 = RQIFYARYVGWFLAFPWPIIQMSLLGG
length = 27, ident = 37%, score = 54.0, nSDMC = 6.5
```

YRO2_YEAST versus YKQ3_CAEEL

```
YRO2_YEAST tmd01 = HITSRGSDWLFTVFCVNLLFGVILVPLMFRK
YKQ3_CAEEL tmd01 = HASSRGNISIFSVFLIPLIAYILILPGVRRK
length = 31%, ident = 35%, score = 57.0, nSDmc = 7.1
YRO2_YEAST tmd07 = FYGIIDLLILSILPVLFMPLANYLGIERLGLIFDEEPAEHVGPVAEK
YKQ3_CAEEL tmd05 = FYLIFAIGILCVLCGLGLGICEHWRIYTLSTFLDASLDEHVGPKWKK
length = 47, ident = 34%, score = 53.0, nSDmc = 6.8
```

within the membrane which then associates into the polytopic tertiary structure (1), rapid mutational divergence may be better tolerated than with soluble proteins. As a result, vestiges of common ancestry rapidly disappear in the process of evolution. Moreover, as suggested by the YRO2-YKQ3 alignment, and our previous data with membrane transporters (19), the genes encoding the primary structure of polytopic membrane proteins may be modular, each module representing one or several TMD segment that can be rearranged, deleted, inserted, or duplicated. These factors combined impede conventional analysis of evolutionary relationships.

To resolve these questions will require the availability of additional sequences that provide the links between structures that are too dissimilar to support a finding of probable homology. To resolve this, and in view of the rapidly increasing sequence database, we will periodically repeat INCA on the same structures analyzed here. Only if the sequence similarities shown here between bacteriorhodopsins and yeast structures and more distant proteins represent true homology and thus common ancestry, would we expect to find these missing evolutionary links between them. The outcome of this analysis could clarify how polytopic membrane proteins have evolved, and help us better understand their structure and function.

ACKNOWLEDGMENTS

Supported in part by research grants form the NIH, DA04166, GM43102, and GM37188.

REFERENCES

1. J.-L. Popot and D. M. Engelman. *Biochem.* **29**:4031–4037 (1990).
2. J. Soppa. *FEBS Let.* **342**:7–11 (1994).
3. E. W. Taylor and A. Agarwal. *FEBS Let.* **325**:161–166 (1993).
4. M. B. Tsendina, D. I. Frishman, V. F. Levchenko, and V. L. Berman. *J. Evolut. Biochem. Physiol.* **34**:600–609 (1989).
5. R. Henderson, J. M. Baldwin, T. A. Ceska, F. Zemlin, and K. H. Downing. *J. Mol. Biol.* **213**:899–929 (1990).
6. L. Pardo, J. A. Ballesteros, R. Osman, and H. Weinstein. *Proc. Natl. Acad. Sci. USA* **89**:4009–4012 (1992).
7. F. X. Schertler, C. Villa, and R. Henderson. *Nature* **362**:770–772 (1993).
8. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. *J. Mol. Biol.* **215**:403–10 (1990).
9. Y. Sugiyama, M. Maeda, M. Futai, and Y. Mukohata. *J. Biol. Chem.* **264**:20859–20862 (1989).
10. G. Aljinovic and T. M. Pohl. *Yeast* **11**:475–479 (1995).
11. R. Wilson, R. Ainscough, K. Anderson, et al. *Nature* **368**:32–38 (1994).
12. D. M. Hillis. *Syst. Biol.* **41**:268–269 (1992).
13. S. Karlin and S. F. Altschul. *Proc. Natl. Acad. Sci. USA* **87**:2264–2268 (1990).
14. S. Karlin and S. F. Altschul. *Proc. Natl. Acad. Sci. USA* **90**:5873–5877 (1993).
15. S. Henikoff and J. G. Henikoff. *Proc. Natl. Acad. Sci. USA* **89**:10915–10919 (1992).
16. S. F. Altschul. *J. Mol. Biol.* **219**:555–565 (1991).
17. J. C. Wootton and S. Federhen. *Comp. Chem.* **17**:149–163 (1993).
18. W. R. Pearson and D. J. Lipman. *Proc. Natl. Acad. Sci. USA* **85**:2444–2448 1988).
19. R. C. Graul and W. Sadée. *Pharm. Res.* **14**:388–400 (1997).
20. M. G. Claros and G. von Heijne. *CABIOS* **10**:685–686 (1994).
21. D. C. Rees, L. DeAntonio and D. Eisenberg, *Science* **245**, 510–513 (1992).
22. R. P. Riek, M. D. Handschumacher, S. S. Sung, M. Tan, M. J. Glynias, M. D. Schluchter, and R. M. Graham. *J. Theoret. Biol.* **172**:245–258 (1995).
23. B. Persson and P. Argos, *J. Mol. Biol.* **237**, 182–192 (1994).
24. M. Regnacq and H. Boucherie, *Curr. Genet.* **23**, 435–442 (1993).