Volume 16 Number 18 1988 Nucleic Acids Research

Conservation of short patches of amino acid sequence amongst proteins with a common function but evolutionarily distinct origins: implications for cloning genes and for structure-function analysis

Juergen K.V.Reichardt and Paul Berg

Department of Biochemistry, Stanford University School of Medicine, Stanford, CA 94305, USA

Received June 20, 1988; Revised and Accepted August 18, 1988

ABSTRACT

Small patches of identical amino acid sequences commonly occur in proteins that have the same function but are derived from evolutionarily distant organisms. Reverse translation of such patches into degenerate pools of oligonucleotides provide useful hybridization probes for cloning the gene for the corresponding protein from other organisms. Since the conserved patches of identical amino acid sequence are probably important for the protein's biological function, they are preferred targets for reverse genetic studies aimed at defining structure-function relationships.

INTRODUCTION

Two general methodologies are widely used to identify genes of interest amongst cloned DNA populations. One relies on the expression of the cloned sequence to yield a product which can be detected either by its immunological reactivity or its biological function. The second approach depends on hybridization with labeled nucleic acid probes to detect the clones of interest. Where amino acid sequences of the encoded protein are known, these can be "reverse-translated" into pools of degenerate oligonucleotides for use as hybridization probes. Thelander and Berg (1) showed that a common amino acid sequence in the M2 subunit of ribonucleotide reductases from Epstein-Barr virus, Herpes simplex type II, clam and E. coli provided an oligonucleotide pool that aided in the cloning of the cDNA encoding the mammalian protein. Recently, we adopted this approach to clone the cDNA encoding the human galactose-1-phosphate uridyl transferase (2). In this instance, highly degenerate oligonucleotide pools whose sequences were based on short patches of amino acid sequence identity between the galactose-1-phosphate uridyl transferases from E. coli and S. cerevisiae were used as hybridization probes to detect the desired cDNA clones (2). Hanks (3) has also shown that genes encoding enzymes that share a related function - phosphorylation of serine residues in proteins - can be cloned using hybridization probes based on amino acid sequences common to this class of enzymes.

In this paper, we note that proteins with identical functions, but derived from evolutionarily distant organisms, commonly share short patches of identical amino acid sequences at corresponding locations throughout the polypeptide chain. These patches pinpoint those parts of the sequence that are most useful for synthesizing oligonucleotide hybridization probes for use in identifying the gene (or cDNA) from other organisms. By contrast, the nucleotide sequences encoding these patches of identical amino acid sequence are rarely identical, making them far less reliable as hybridization probes for related genes. We suspect that such patches are a reflection of their essential roles in the protein's biological function. Focussing mutational modifications to the nucleotide sequences that encode these patches should aid in studies attempting to correlate structure and function.

METHODS

The amino acid sequences of the proteins described here were found using release 12.0 of the protein (PIR) database on the Bionet system. Enzymes whose amino acid sequence were known for at least three distantly related organisms (e.g. E. coli, S. cerevisiae and a higher eukaryote) were examined for the existence of short patches of amino acid identity (4 to 6 residues long) using an algorithm written by Peter Griffin (Stanford University) that runs on an Apple Macintosh personal computer. Protein sequences were aligned using Bionet's GENALIGN multiple sequence alignment program.

RESULTS AND DISCUSSION

Much of the modest overall homology (40%) between the amino acid sequences of the galactose-1-phosphate uridyl transferases (GalT) of E. coli and S. cerevisiae resides in relatively short patches of 5 or 6 identical amino acids, many of which are also conserved in the corresponding human protein (2). To test the proposition that such patch identities are a general feature of amino acid sequences in proteins with the same function but from evolutionarily distant organisms, we aligned and compared the amino acid sequences of eleven proteins generally classified as housekeeping enzymes (Table I). These sequences were selected from a list of roughly 40 enzymes because their sequences were known from at least three distantly related organisms (e.g. a bacterium, a lower and a higher eukaryote). Ten of the 11 proteins had one or more sequence patches with 5 or more identical amino acids, located at approximately the same positions in the functionally related proteins. The one apparent exception, citrate synthase, contains one patch of

TABLE 1
SEQUENCE CONSERVATION OVER LONG EVOLUTIONARY DISTANCES

ENZYME	NUMBER OF COMMON SEQUENCES	
	5 amino acids	6 or more amino acids
Alcohol Dehydrogenase (yeast, maize, human)	2	1
Citrate Synthase* (*E.coli*, yeast, human)	0	0
Cytochrome *c* Oxidase II (yeast, maize, bovine)	3	1
Dihydrofolate Reductase (*E.coli*, yeast, human)	1	5
Galactose-1-phosphate Uridyl Transferase (*E.coli*, yeast, human)	4	0
Glyceraldehyde-3-phosphate Dehydrogenase (*B.stearothermophilus*, yeast, human)	4	3
Ribonucleotide Reductase subunit 2 (*E.coli*, yeast, human)	1	0
6-Phosphofructokinase (*E.coli, B.stearothermophilus*, rabbit)	1	0
Pyruvate Kinase (*E.coli*, yeast, chicken)	3	1
Triosephosphate Isomerase (*B.stearothermophilus*, yeast, human)	0	3
Superoxide Dismutase (*E.coli*, yeast, human)	8	4

*: Note that while there are no patches 5 or more amino acids long for all three citrate synthases, there is one patch 4 amino acids long and many regions which are clearly homologous if conservative changes are allowed.

4 identical amino acids, but as is the case for the larger patches in the other proteins, this single patch is flanked by mismatches of structurally related amino acids. The nature and location of the patch homologies uncovered by a comparison of three triose-phosphate isomerases and of three glyceraldehyde-3-phosphate dehydrogenases are shown in Figure 1. Based on this relatively limited set of comparisons, we surmise that the occurrence of such patches of sequence homology are probably a general feature of proteins with a common function in widely divergent species. Most likely, such sequence conservation is a consequence of their contribution to the protein's function.

The existence of such patch identities is more easily recognized by comparisons between the amino acid sequences of proteins from widely divergent rather than closely related species. Thus, a comparison of the amino acid

A

GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE

```
yeast       VRVAINGFGR GRLVMRIALQRKNVEVVALNDPFISNDYSAYMF KYS
human     MGKVKVGVNGFGRIGRLVTRAAFNSGKVDIVAINDPFIDLNYMVYMFQYDS
B.stearo.   AVKVGINGFGRIGRNVFRAALKNPDIEVVAVND LTNADGLAHLLKYDS

consensus ---V-V--NGFGR-GR-V-R-AL-------VA-ND---------------S


          THGRYAGEVSHDDKHIIVDGHKIATFQERDP ANLPWASLNIDIAISTGVF
          THGKFHGTVKAENGKLVINGNPITIFQERDPSKIKWGDAGAEYVVESTGVF
          VHGRLDAEVVVNDGDVSVNGKEIIVKAERNPENLAWGCIGVDIVVESTGRF

          -HG-----V----------G--I----ER-P---------------STG-F


          KELDTAQKHIDAGAKKVVITAP SSTAPMFVMGVNEEKY TSDLKIVSNAS
          TTMEKAGAHLQGGAKRVIISAP SADAPMFVMGVNHEKYD NGLKIISNAS
          TKREDAAKHLEAGAKKVIISAPAKVENITVVMGVNQDKYDPKAHHVISNAS

          -----A--H---GAK-V-I-AP--------VMGVN--KY--------SNAS


          CTTNCLAPLAKVINDAFGIEEGLMTTVHSMTATQKTVDGPSHKDWRGGRTA
          CTTNCLAPLAKVIHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGA
          CTTNCLAPFAKVLHQEFGIVRGMMTTVH SYTNNQRILDLPHKDLRGARAA

          CTTNCLAP-AKV----FGI--G-MTTVH--------------K--R--R-A


          SGNIIPSSTGAAKAVGKVLPELQGKLTGMAFR PTVDVSVVBLTVKLNKET
          LQNIIPASTGAAKAVGKVIPELDGKLTGMAFRVPTANVSVVDLTCRLEKPA
          AESIIPTTTGAAKAVALVLPELKGKLNGMAMRVPTPNVSVVDLVAELEKEV

          ---IIP--TGAAKAV--V-PEL-GKL-GMA-R-PT--VSVV-L---L-K--


          TYDEIKKVVKAAAEGKLKGVLGYTEDAVVSSDF LGSNSSIFDAAAGIQLS
          KYDDIKKVVKQASEGPLKGILGYTEHQVVSSDFNSDTHSSTFDAGAGIALN
          TVCEVNAALKAAAEGELKGILAYSEEPLVSRNYNGSTVSSTIDALSTMVID

          ---------K-A-EG-LKG-L-Y-E---VS--------SS--DA-------


          PKFVKLVSWYDNEYGYSTRVVDLVEHVAKA
          DHFVKLISWYDNEFGYSNRVVDLMAHMASKC
          GKMVKVVSWYDNETGYSHRVVDLAAYINAKGL

          ---VK--SWYDNE-GYS-RVVDL---------
```

B

TRIOSE-PHOSPHATE ISOMERASE

```
yeast       MARTFFVGGNFKLNGSKQSIKEIVERLNTASIPENVEVVICOPATYLDYSV
human      APSRKFFGVGNWKMNGRKQSLGELIGTLNAAKVDADTEVVCAPPTAYID FA
B.stearo.    MRKPIIAGNWKMHKTLAEAVQFVEDVKGHVPPADEVDSVVAP FLFRLVQ

consensus  --R-----GN-K-----------------------------P--------

           SLVKKPQVTVGAQNAYLKASGAFTGENSVDQIKDVGAKWVILGHSERRSYFH
           RQKLDPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATWVVLGHSERRHVFG
           AADGTDLQKIGAQTMHFABZGAYTGEVSPVMLKDLGVTYVILGHSERRHMFA

           -----------AQ-------GA-TGE-S----KD-G---V-LGHSERR--F-

           EDDKFIADKTKFALGQGVGVILCIGETLEEKKAGKTLDVVERQLNAVLEEVK
           ESDELIGQKVAHALAEGLGVIACIGEKLDEREAGITEKVVFEQTKVIADNVK
           ZTBZTVBKKVLAAFTRGLIPIICCGESLEERQAGETDAVVSQVEKALAGLTP

           --------K---A---G-----C-GE-L-E--AG-T--VV------------

           DWTNVVVAYEPVWAIGTGLAATPEDAQDIHASIRKFLASKLGDKAASELRIL
           DWSKVVLAYEPVWAIGTGKTATPQQAQEVHEKLRGWLKSNVSDAVAQSTRII
           QEVKIILAYEPLWAIGTGKSSTPZBABSVCGHIRSVVSRLFGPEAAEAIRIQ
                    *
           -------AYEP-WAIGTG----P----------R-----------A------

           YGGSANGSNAVTFKDKADVDGFLVGGASLKPEFVDIINSRN
           YGGSVTGATCKELASQPDVDGFLVGGASLKPEFVDIINAKQ
           YGGSVKPDNIRDFLAZZZIBGALVGGASLEPASFLQLVQAGRHE

           YGGS----------------G-LVGGASL-P-------------
```

Figure 1: Sequences from the same protein from distantly related organisms show only patch identities.

Protein sequences of the same enzyme from three distantly related organisms (a prokaryote, a lower and a higher eukaryote) were aligned with the BIONET GENALIGN multiple sequence alignment program. Underlined are those patches that are 5 or more amino acids in length.

sequences in the human and rabbit triose-phosphate isomerases, which are 98% identical, shows extended rather than patch identities (Figure 2, consensus^{M}). When the comparison also includes the enzymes from the chicken and a vertebrate fish, which are 80 to 90% identical to the mammalian enzymes, such identical patches become more evident (Figure 2, consensus^{V}). Extending the comparison to include the S. cerevisiae (Figure 2, consensus^{E}) and B. stearothermophilus enzymes (Figure 2, consensus^{T}), each of which is 40 to 50%

TRIOSE-PHOSPHATE ISOMERASE

```
consensus^m   APSRKFFVGGNWKMNGRK--LGELI-TLNAAKVPADTEVVCAPPTAYID
consensus^V   AP-RKFFVGGNWKMNG----LGELI-TL--AK-----E-VC--P--Y-D
consensus^E   ---R-FFVGGN-K-NG------E----L--A------E-V---P--Y-D

CONSENSUS^T   ---R-----GN-K------------------------------P-----

              FARQKLDPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATWVVLGHSERRHV
              FAR-K-D-K--VAAQNCYKV--GAFTGEISP-MIKD-G--WV-LGHSERRHV
              -----------V-AQN-Y----GAFTGE-S---IKD-G--WV-LGHSERR--

              -------------AQ-------GA-TGE-S----KD-G---V-LGHSERR--


              FGESDELIGQKVAHAL EGLGVIACIGEKLDEREAGITEKVVFEQTKVIADN
              FGESDELIGQKV-HAL-EGLGV-ACIGEKLDEREAGITE-VVF--T--IAD-
              F-E-D--I--K---AL--G-GV--CIGE-L-E--AG-T--VV----------

              ----------K---A---G-----C-GE-L-E--AG-T--VV----------


              VKDWSKVVLAYEPVWAIGTGKTATPQQAQEVHEKLRGWLKSNVSDAVAQSTR
              VKDWSKVVLAYEPVWAIGTGKTA-PQQ-QE-H-KLR-WLK--VS--VA---R
              VKDW--VV-AYEPVWAIGTG--A-P---Q--H---R--L--------A---R
                       *
              ---------AYEP-WAIGTG----P----------R----------A----


              IIYGGSVTGATCKELASQPDVDGFLVGGASLKPEFVDIINAKQ
              IIYGGSVTG--CKELAS--DVDGFLVGGASLKPEFV-------
              I-YGGS--G----------DVDGFLVGGASLKPEFV-------

              --YGGS---------------G-LVGGASL-P-------------
```

Figure 2: Patch identities are evident only in distantly related organisms. Sequences encoding triose-phosphate isomerase from different organisms were aligned. Consensus^M is the sequence shared by mammals (human and rabbit), consensus^V applies to the vertebrate enzymes (mammals, chicken and a fish, Latimeria), consensus^E summarizes the identical residues amongst all eukaryotes (consensus^V plus yeast) and consensus^T is the overall consensus for all species (including B. stearothermophilus).

homologous to the vertebrate enzymes, reduces the number of patches with five or more identical amino acids to 3. Thus, comparisons between amino acid sequences amongst proteins from the most divergent species, but with a common function are most likely to be helpful for the purposes described below.

Patch identities of the type shown in Figure 1 have several practical applications. One is to facilitate the cloning of protein coding genes from

any species if the amino acid sequences of the corresponding proteins from more evolutionarily distant species are known or can be deduced. Thus, degenerate pools of oligonucleotides derived by reverse translation of the shared amino acid sequences in the reference proteins provide suitable hybridization probes for screening libraries of cloned genes or cDNAs (1,2,3). For this purpose, the patches must be 5 amino acids or larger because oligonucleotides shorter than 15 nucleotides are likely to give spurious signals with unrelated sequences. The greatest sensitivity and selectivity is achieved from patches that are reverse translatable into the fewest possible oligonucleotide coding sequences. Also, having two or more different oligonucleotide probes is most useful since a sequence of 5 amino acids might by chance occur in proteins other than the one of interest. During the cloning of the human GALT cDNA (2) we, in fact, isolated 21 clones. Only one of them hybridized to all six oligonucleotides tested and none of the other cDNAs hybridized to more than one probe.

Pools of degenerate oligonucleotide sequences produced by reverse translation of amino acid sequences are more likely to detect the unknown gene than are the actual nucleotide sequences that encode the common patch. Because of the degeneracy in the genetic code and preferred codon uses amongst the different species, the nucleotide sequences corresponding to the same amino acid sequence in four patches of the E. coli, S. cerevisiae and human GalT proteins vary at multiple positions (Figure 3A, B, D, E). These differences preclude the use of nucleic acids that encode the protein in the reference sequences as probes to identify the unknown gene. Figure 3C shows a patch where the human sequence differs from the corresponding yeast and E. coli consensus sequences by two amino acids within the patch. These mismatches are conservative changes at the structural level. Note that the discordances between the nucleotide sequence encoding the non-matching amino acid sequences resemble the extent of mismatches between the same sequences.

Patch identities amongst proteins with the same or closely related functions can also serve to direct attention to residues or regions that may be relevant to the proteins' common function. In attempting to identify the critical histidine residue to which the uridyl moiety of uridine-diphosphoglucose becomes linked during the enzymatic reaction, Frey et al. (6) substituted an asparagine residue for each histidine in the E. coli enzyme and assayed the catalytic activities of the modified proteins. Only the histidines at positions 164 and 166 were essential. These two histidines are the only ones of fifteen histidines that are conserved in the GALT enzymes

GALACTOSE-1-PHOSPHATE URIDYL TRANSFERASE

A) protein sequence:

consensus yeast-E.coli	PHE GLU ASN LYS GLY ALA
human	PHE GLU ASN LYS GLY ALA

DNA sequence:

E.coli sequence	TTT GAA AAC AAA GGC GCG
yeast sequence	TTT GAA AAC AAA GGT GCC
human sequence	TTT GAA AAC AAA GGT GCC

B)

consensus yeast-E.coli	ala MET GLY CYS SER ASN
human	met MET GLY CYS SER ASN
E.coli sequence	GCG ATG GGC TGC TCT AAC
yeast sequence	GCC ATG GGT TGT TCC AAC
human sequence	ATG ATG GGC TGT TCT AAC

C)

consensus yeast-E.coli	TRP PRO phe glu THR LEU
human	TRP PRO tyr gln THR LEU
E.coli sequence	TGG CCG TTC GAA ACG CTA
yeast sequence	TGG CCA TTT GAG ACC TTG
human sequence	TGG CCC TAC CAG ACA CTG

D)

consensus yeast-E.coli	TYR ASP ASN LEU PHE
human	TYR ASP ASN LEU PHE
E.coli sequence	TAT GAC AAC CTC TTC
yeast sequence	TAT GAT AAT TTA TTT
human sequence	TAT GAC AAC CTC TTT

E)

consensus yeast-E.coli	GLU GLN ALA ALA GLU
human	GLU GLN ALA ALA GLU
E.coli sequence	GAA CAG GCA GCA GAG
yeast sequence	GAA CAA GCT GCT GAA
human sequence	GAG CAG GCT GCA GAG

Figure 3: Patch identities are DNA sequences unsuitable as hybridization probes.

The DNA sequences encoding five conserved patches from the E. coli, yeast and human GalT proteins are compared. The mismatches are underlined. Totally degenerate oligonucleotide pools would have matched each sequence.

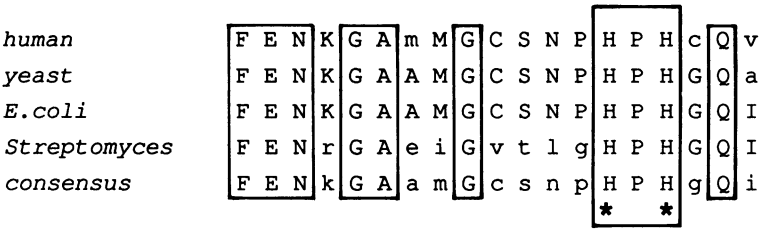

Figure 4: The active site histidines of GalT.
The only two conserved histidines of this enzyme are part of a conserved patch. The two histidines that are essential for the enzyme's activity in E. coli occur in one of the amino acid sequence patches common to the E. coli, S. cerevisiae, S. lividans (12) and human proteins.

from four distantly related species (Fig. 4). Indeed, the sequences surrounding the common his-pro-his is nearly identical in the human, E. coli and yeast enzymes. A comparison of the four GalT proteins also revealed a shared cysteine residue surrounded by other conserved amino acid residues; perhaps the sensitivity of GalT to inactivation by sulfhydryl reagents (7) is due to this cysteine residue (Human cys 128, E. coli cys 108).

The inference that the amino acids comprising the conserved patch identities are important for the protein's functions is supported by an examination of the catalytic or co-factor binding domains of other enzymes. Thus, amongst the short patches that are identical amongst the glyceraldehyde-3-phosphate dehydrogenases of the three species shown (Fig. 1A), four are part of the NAD binding site and 3 patches located in the middle of the chain form part of the catalytic center (4). Similar correspondences between the conserved patches and the domains responsible for substrate binding or catalytic activity were amongst the various alcohol dehydrogenases and triosephosphate isomerases (5).

Conserved amino acid sequence patches in enzymes having the same or closely related functions, but derived from evolutionary distant species, have also been noted by others. Patthy (8) showed that patches of amino acid sequence identities amongst distantly related proteins could be used to identify new members of two protein families even though there was only marginally significant sequence homologies outside the patches. A comparison of the amino acid sequences of E. coli malate dehydrogenase with the sequences of the corresponding enzymes from yeast and pig mitochondria revealed that although their sequences are only about 48 to 58% identical, they share six conserved patches of five or more identical amino acids (9). These conserved regions contain residues and structural elements important for NAD(H)-binding, for

formation of the subunit interface, and for catalysis. The absence of these patches in the pig's cytoplasmic enzyme led McAlister-Henn (9) to argue in favor of the prokaryotic origin of eukaryotic mitochondrial malate dehydrogenases. Clarke et al. (10) identified regions of strong sequence identity (stretches of at least ten amino acids with 60% or more colinear amino acid identity) in the isoleucyl-tRNA synthetases of S. cerevisiae and E. coli. Because the corresponding region in the B. stearothermophilus tyrosyl-tRNA synthetase is involved in tyrosine binding, they focussed their mutagenesis on this patch in the E. coli isoleucyl-tRNA synthetase. They found that certain substitutions within the patch caused reduced isoleucine binding. Conserved patches amongst various prokaryotic and eukaryotic biotin-dependent carboxylases have also been identified by Samols et al. (11). One such patch of conserved amino acids surrounded the biotinyl lysine that is involved in the catalytic reaction common to this group of enzymes. Other patches of conserved amino acids amongst the carboxylases and decarboxylases have been implicated in other common steps in the reactions catalyzed by these enzymes (11).

ACKNOWLEDGMENTS

This work was supported by Public Health Service grant GM-13235 from the National Institute of General Medical Sciences. JR thanks Peter Griffin (Stanford University) for help with the database searches and Janette Carey and Steve Lee for helpful discussions.

REFERENCES

1. Thelander, L. and P. Berg (1986) Molec. Cell Biol. 6, 3433-3442.
2. Reichardt, J.K.V. and P. Berg (1988) Mol. Biol. Med. 5, 107-122.
3. Hanks, S.K. (1987) Proc. Natl. Acad. Sci. USA 84, 388-392.
4. Biesecker, G., J. J. Harris, J. C. Thierry, J. E. Walker and A.J. Wonacott (1977) Nature 266, 328-33.
5. Banner, D.W., A. C. Bloomer, G.A. Petsko, D.C. Philips, C.I. Pogson, I.A. Wilson, P.H. Corran, A.J. Furth, J.D. Milman, R.E. Offord, J.D. Priddle, S.G. Waley (1975) Nature 255, 609-614.
6. Frey, P.A., personal communication.
7. Williams, V. P., C. Fried and G. Popjak (1981) Biochem. Biophys. Res. Comm. 206, 353-361.
8. Patthy, L. (1987) J. Molec. Biol. 198, 567-577.
9. McAllister-Henn, L. (1988) Trends Biochem. Sci. 13, 178-181.
10. Clarke, N. D., D. C. Lien and P. Schimmel (1988) Science 240, 521-523.
11. Samols, D., C. G. Thornton, V. L. Murtif, G. K. Kumar, F. C. Haase and H. G. Wood (1988) J. Biol. Chem. 263, 6461-6464.
12. Adams, C. W., J. A. Fornwald, F. J. Schmidt, M. Rosenberg and M. E. Brawner (1988) J. Bacteriol. 170, 203-212.