Volume 16 Number 13 1988 Nucleic Acids Research

Detection of protein similarities using nucleotide sequence databases

Steven Henikoff* and James C.Wallace

Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Seattle, WA 98104, USA

Received March 8, 1988; Revised and Accepted May 16, 1988

ABSTRACT

A simple procedure is described for finding similarities between proteins using nucleotide sequence databases. The approach is illustrated by several examples of previously unknown correspondences with important biological implications: Drosophila elongation factor Tu is shown to be encoded by two genes that are differently expressed during development; a cluster of three Drosophila genes likely encode maltases; a flesh-fly fat body protein resembles the hypothesized Drosophila alcohol dehydrogenase ancestral protein; an unknown protein encoded at the multifunctional E. coli hisT locus resembles aspartate β-semialdehyde dehydrogenase; and the E. coli tyrR protein is related to nitrogen regulatory proteins. These and other matches were discovered using a personal computer of the type available in most laboratories collecting DNA sequence data. As relatively few sequences were sampled to find these matches, it is likely that much of the existing data has not been adequately examined.

INTRODUCTION

The detection of amino acid sequence similarities between proteins identified in unrelated studies has proven to be a valuable tool for studying gene function. Important examples include the similarities between various oncogenes and proteins of known function, first with the finding that the v-sis protein is related to platelet derived growth factor (1), and more recently with the finding that the jun protein partially resembles yeast GCN4, a DNA-binding regulatory protein (2). Such discoveries were possible because amino acid sequences of known proteins have been searched for unexpected similarities to new sequences. The protein databases that have been used for such searches (3) are neither complete nor up-to-date, since the large majority of new sequence data comes from nucleotide sequence, initially collected in DNA databases. Therefore, it has been difficult for researchers to do an exhaustive survey of all sequences for similarities to newly determined ones.

We show here that searching of DNA databases for amino acid sequence similarities is an effective and reliable approach to finding useful

correspondences between proteins. We illustrate this approach with several examples of amino acid sequences of previously unknown function which we show are related to other amino acid sequences, leading to unexpected biological insights in each case. This simple method was carried out entirely on a relatively modest personal computer using a commercially available software package; therefore, the approach is available to the general population of molecular biologists and biochemists. The ease of finding previously unrecognized matches, several of which are of clear importance, indicates that existing data often are not adequately analyzed.

MATERIALS AND METHODS

Database searches and dot matrix and alignment procedures were performed using GENEPRO version 4.1 software and GenBank 52.0 obtained from Riverside Scientific (18332 57th Ave. N.E., Seattle, Washington 98155). Searches were done on a BIOS AT personal computer (IBM AT-compatible) equipped with a Miniscribe 44 Meg hard disk, obtained from Lang Systems, Arlington, Massachusetts.

Generally, a "query sequence" or "probe", consisting of an unidentified open reading frame (ORF) obtained from the GenBank 52.0 database, was conceptually translated into protein and compared to every possible translated reading frame of every nucleotide sequence in the database. The software that we used performed these operations by fetching an individual nucleotide sequence, translating each reading frame into protein, comparing that reading frame with the probe, repeating the comparison for the next reading frame and then repeating the entire operation for the next nucleotide sequence.

The comparison strategy was to align 30 amino acids of the probe with a stretch of 30 amino acids from a database sequence and calculate a log-odds score, which measures the likelihood that two aligned amino acids are functionally equivalent (4, 5). The probe was then aligned with the next stretch of translated sequence and a log-odds score calculated. In order to reduce the number of unproductive alignments, only sequences in which one or more dipeptides match between the probe and the translated database sequences were considered (6). To reduce the likelihood of missing optimal matches, all possible ungapped alignments including one or more dipeptide matches were made (7). The entire procedure typically required a few hours to search the entire GenBank 52.0 database on our "AT clone" computer. Although we used the search comparison procedure in the GENEPRO software package, the DNA database searching strategy should be easily incorporated into other search programs.

Following each completed search, the distribution of the best log-odds

score for each DNA sequence was inspected. Typically, the best log-odds scores for spurious matches were 9.8 using a window of 30 residues. For each of the matches described in this report, a score of at least 10.4 was obtained in the search. Standard dot matrix and alignment procedures based on log-odds scores were then performed on the two protein sequences in order to extend and verify the similarity. In each case reported here, a match obtained by database searching was found to extend much further, particularly when occasional gaps were introduced.

RESULTS

Searching of DNA databases for amino acid sequence similarities

The searching strategy involves comparison of a probe to an entire DNA database translated into all possible reading frames. A standard searching protocol for protein similarities is used. The effectiveness of this procedure is demonstrated by examination of the log-odds scores for a search of GenBank 52.0 using a portion of the Drosophila melanogaster Antp gene predicted amino acid sequence as probe (8, 9). A conserved 61 amino acid segment of this sequence, called the "homeo-box", has been found in this and several other Drosophila genes that regulate segmentation in the fly, leading to the isolation of homologous segments from unrelated organisms. When the homeo-box amino acid sequence is used to probe GenBank, a clearly biphasic log-odds distribution is seen, with 37 entries showing best log-odds scores between 10.4 and 14.0, while the remaining ca. 14,000 entries showed log-odds scores of 9.8 or less. All 37 entries have been identified as homeo-box proteins, mostly isolated on that basis. A search of the GenBank 52.0 annotations fails to find a known homeo-box protein that is not among these 37 entries, indicating that this search procedure is capable of picking up all real homologies without interference from spurious matches.

The following examples illustrate the types of correspondences that have been found using this procedure for a variety of probe sequences. Each of the 10 homologies described below does not appear to have been previously recognized, as determined by searches of the literature and by contact of experts in the appropriate areas.

Two differently regulated Drosophila genes encode elongation factor Tu

In a search for genes expressed specifically in females, two unlinked and abundantly expressed genes were found to encode proteins that are identical at 90% of aligned residues (10). The mRNA of one of these genes, called F1, accumulates at highest levels early in embryogenesis, with gradual reduction in amount during development. Female adults show very high transcript levels.

```
Dro  MGKEKIHINIVVIGHVDSGKSTTTGHLIYKCGGIDKRTIEKFEKEAQEMGKGSFKYAWVLDKLKAERERGITIDIALWKFETAKYYVTIIDAPGHRDFIK  100
Art
Hum       T                                   A                              S     S
Yea       S   V                               A L                                  P  Q  V
Muc       T V V                           E   A L                                  P  N  V

Dro  NMITGTSQADCAVQIDAAGTGEFEAGISKNDQTREHALLAFTLGVKQLIVGVNKMDSSEPPYSEARYEEIKKEVSSYIKKVGYNPAAVAFVPISGWHGDN  200
Art             L V  V          G          Y              T   F    F        A   I
Hum             L V  V          G          Y              T     QK     V    T   I    DT         N
Yea            IL I G V         DG             R    A     VK--WD S FQ  V  T NF        KT P       N
Muc            IL I G           DG            FR    AI    TTK--W QG  N  V    GF   I G  KS P

Dro  MLEPSTNMPWFKGWEVGRKEGNADGKTLVDALDAILPPARPTDKALRLPLQDVYKIGGIGTVPVGRVETGVLKPGTVVVFAPANITTEVKSVEMHHEALQ  300
Art     A DRL  Y   NIE    K      L         S   E P                          II   MI T                    S E
Hum       A        K T  D   S T  LE   C     T     P                              M  T   V V               S
Yea   I AT  A  Y    KET A VVK     LE I    EQ S     P                            I   M  T    GV             Q E
Muc    DE           NKET A SKT    LE I    E  V  S  P                          TI A M  N    AV             T T

Dro  EAVPGDNVGFNVKNVSVKELRRGYVAGDSKANPPKGAADFTAQVIVLNHPGQIANGYTPVLDCHTAHIACKFAEILEKVDRRSGKTTEENPKFIKSGDAA  400
Art  Q S                    S   N  AR SQ  F          S                      K  C   T     AE
Hum    L               D    N      ND  ME  G      I      SA  A                LK  I      KL DG   L
Yea  QG                     N C  A ND    C S N T          SA  S             R D L   N      KL DH   L
Muc    GL              DI   N CS   ND A ES S       I      SA  A               S LI  I      KM DS   V    S

Dro  IVNLVPSKPLCVEAFQEFPPLGRFAVRDMRQTVAVGVIKAVNFKDASGGKVTKAAEKATKGKK                                    463
Art  MIT             SD                        S     PTA          G K -
Hum    DM  G  M   S SDY                          DK A GA     S Q  Q A -
Yea  L KP     M     S Y                        S D- TEKAA      Q  A K--
Muc    KM     M    YTDY                          E-IVKKA      A  S K--
```

1. D. melanogaster F1 protein (Dr) aligned with cytoplasmic EF-Tu sequences from Artemia salina (Ar), humans (Hu), S. cerevisiae (Ye) and Mucor racemosa (Mu). Only differences from the Drosophila sequence are shown, where dashes indicate missing residues.

The other gene, called F2, appears to be abundantly expressed only during the pupal stage, and is present in both male and female adults. The much higher level of similarity between F1 and F2 at the amino acid sequence level than at the nucleotide sequence level suggested to the authors that the two genes have been maintained by strong selection for biological function. Our search of GenBank using the predicted F1 protein as probe reveals that it, and likely F2 as well, are the D. melanogaster cytoplasmic elongation factors Tu (EF-Tu, sometimes known as EF-1α). This conclusion is based on the striking similarity of F1 to the amino acid sequences of known cytoplasmic EF-Tu proteins (Figure 1). The predicted amino acid sequence of one of these genes, F1, is identical to that of Artemia (brine shrimp) EF-Tu (11) at 88% of the 463 residues. The F1 protein shows 84% identity to human EF-Tu (12), 78% to Saccharomyces cerevisiae EF-Tu (13, 14) and 77% to EF-Tu of Mucor racemosa (15), a mycelial fungus.

EF-Tu is one of the most abundant cellular proteins, promoting the GTP-dependent binding of aminoacyl-tRNA to the ribosome. Developmental regulation of an EF-Tu gene has been recently reported in mouse erythroleukemia cells, where differentiation was found to be correlated with reduced mRNA

```
H      MRPQSAACLLLAIVGFVGAT-EWWESGNYYQIYPRSFRDSDGDGIGDLNGVTEKLQYLKDIGFTGTWLSPIFKSPMVDFGYDISDFYQIHPEYGTM   95
D    MPKWAHLGLAALLLISTTQEGTADIDWWENASLYQIYPRSFQDSDGDGIGDLKGITSRLGYLKEIGITATWLSPIFTSPMSDFGYDISNFYDIDPIFGTL  100
L      MFKLLVLSCLLALALPSLAEVGWWKTGQFYQIYPRSFKDSDGDGVGDLIGITQQLPYLKEIGITATWLSPIFTSPMAKFGYDVAKLKGIDPIFGTM   96
                    *      *  *** *  ***** *****   ** ************ ** *  *  *  *    **  * ******     * ***
Y                  MTISDHPETEPKWWKEATIYQIYPASFKDSNNDGWGDLKGITSKLQYIKDLGVDAIWVCPFYDSPQQDMGYDISNYEKVWPTYGTN   86

H    EDFERMIAKAKEVGIKIILDFVPNHSSTENEWFTKSVD-SDPVYKDFYIWHDGKINNETGEREPPSNWNSEFRYSAWEWNEVRQQYYLHQFAIQQADLNY  194
D    EDFDDLIVEAKSLGVKIILDFVPNHSSDENVWFEKSVN-REDGYDDFYVWDDGKLNEETGARDPPSNWVSVFSGPMWTWNEKRQQYFLHQFQVKQPDLNF  199
L    EDFDDLLARAKELDIKIILDFVPNHTSDECDWFIRSAA-GEEETKDFYVWPTGKVVN--GIRQPPTNWVSVFRGSMWTWNEERQAYYLHQFHAKQPDLNY  193
     **   ** *   ** * * * * ** *** *** *         *   *   *     *  ** ** * * *****  *     **  **  * ***
Y    EDCFELIDKTHKLGMKFITDLVINHCSTEHEWFKESRSSKTNPKRDWFFWRPPKGYDAEGKPIPPNNWKSFFGGSAWTFDETTNEFYLRLFASRQVDLNW  186

H    RNPAVVNEMKN-VIRFWLGKGVSGFRIDAVPYLFEVDLDRYNQYPDEPLTNDSVNCPDPDDHCYTQHIYTQDMPETIDMVYQWRELVDEFHVENGGDKRL  293
D    TNPMVREHMLD-VLKFWLDRGVDGFRIDAVPHIYEHR-NADGSYPDEPV---SGWGSDPNAYDYHDHIYTKDQPATVDLMYEWREFLDNYRAQNGGDSRV  294
L    RNPKVVEAMKD-VLRFWLRKGAYGFRIDAVPHVYEIPADADGNWPDEPR---NEAVSDPEDYTYLQHIYTTDQPETLELVYAFRDVIEEIDAELGGDDRV  289
      *  * *      **** ********    ** *    * *    *                 * * *                 *        * *
Y    ENEDCRRAIFESAVGFWLDHGVDGFRIDTAG-LYSKRPGLPDS-PIFDKTSKLQHPNWGSHNGPRIHEYHQELHRFMKNRVKDGREIMRVGEVAHGSDNA  284

H    LMTEAYTSFENIMTYYGNGVRNGSHIPFNFDFLTSINNASKAGEYVKHIKKWMDASPEGVYANWVLGNHDNKRVASRFGVQRTDLINILLQTLPGHAVTY  393
D    LLAEAYSSVETLSAYFGNSTHQGTQLPMNFQLMYLSGYSTAKDVVGSIDYWMNTMWKEHQTANWVVGNHDTNRVADRMGAHKVDLLNVIVNALPGASVTY  394
L    LLTEAYSPLEVLMQYYGNGTHLGSQIPFNFELLAQISYSSDAYHYSELIGNWLDNMPEGQVANWVFGNHDQSRIGSRLGADRIDACNMIILGLPGVSVTY  389
     * * *   **             **  *           *    *      * *   *   **** *  ** *             *
Y    LYTSAARY-EVSEVFSFTHVEVGTSPFFRYNIVPFTLKQWKEAIASNFLFINGTDSWATTYIE----NHDQARSITRFADDSPKYRKISGKLLTLLECSL  379

H    NGEELGMTDVWISWEDTVDPNACNSDPDNYYARSRDPARSPYQWDASSKAGFTSADHTWLPVADDYKTNNALQQLRAPRSHLQIFKKLXRVRKEPSFRQG  493
D    YGEEIGMSNVDV--ECTGD-----------SCEDRDGERTPMQWTAGKNADFSDGESTWLPLSPEYQRYNVQTERGVSRSSLNIFKGLQELKSSSAFLAF  481
L    QGEEMGMTDVWISWEDTVDPQACQSNEQEFERLTRDPVRTPFQWSDEVNAGFSNASVTWLPVASNYKLVNVKKERGIALSHLNVYKQLRALRDEPTLKQG  489
      *         *                         *    * * * *          *        *    *       *    *   ** **
Y    TGTLYVYQGQEIGQINFKEWPIEKYEDVDVKNNYEIIKKSFGKNSKEMKDFFKGIALLSRDHSRTPMPWTKDK--PNAGFTGPDVKPWFFLN--ESFEQG  475

H    ELNIQAIDDDVIITSRQVYQSLRSLSFDY                                                                        522
D    KEDGGFSYEAVTEQVLQIIRYVKQILF                                                                          508
L    DVSVTAIGPNVLAFKR                                                                                     505
         * * **
Y    INVEQESRDDDSVLNFWKRALQARKKYKELMIYGYDFQFIDLDSDQIFSFTKEYEDKTLFAALNFSGEEIEFSLPREGASLSFILGNYDDTDVSSRVLKP  575

Y    WEGRIYLVK                                                                                            584
```

2. D. melanogaster predicted H, D, and L proteins aligned with one another and with S. carlsbergensis (Y) maltase. Asterisks indicate identical residues between yeast maltase and any one of the Drosophila sequences.

accumulation (16). It was suggested that changes in abundance of EF-Tu might lead to the preferred synthesis of certain proteins. If so, then complex in vivo expression of these two "housekeeping" genes in Drosophila is potentially responsible for translational regulation of large numbers of genes involved in development. Our finding is expected to lead to further studies of EF-Tu in Drosophila, likely the ideal organism for developmental studies.

A cluster of three similar Drosophila genes appear to encode maltases

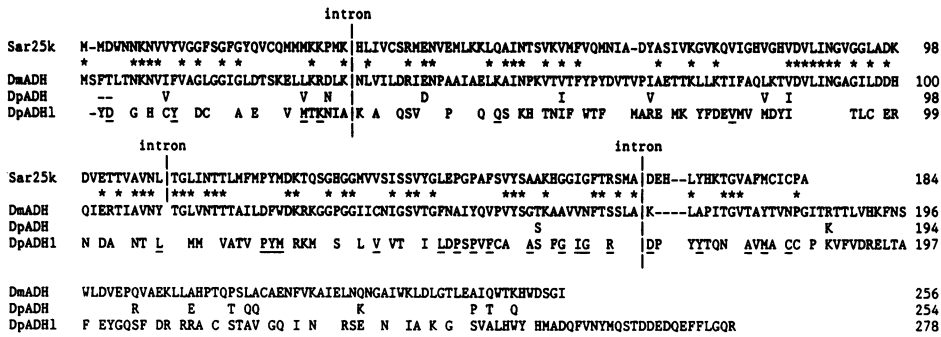
In order to determine the limits of a cluster of larval cuticle protein genes in D. melanogaster, Snyder and Davidson (17) characterized an immediately adjacent region in molecular detail. This region, called HDL, itself encoded a cluster of three related genes on both strands. Sequence and transcriptional analyses of H, D and L suggested that all encoded secreted proteins of similar function needed during both larval and adult stages. Our search of GenBank reveals that the HDL proteins are strikingly similar to the

amino acid sequence predicted for yeast maltase, the enzyme that hydrolyzes maltose to two molecules of glucose (18). S. carlesbergensis maltase aligns with the HDL proteins revealing a level of amino acid sequence similarity of 33% overall, much higher in the NH_2-terminal one-third of the molecules (Figure 2). The HDL proteins are not likely to be enzymes other than maltase, since several other disaccharidases have been sequenced, yet none show significant similarity to yeast maltase or to the HDL proteins (data not shown). It is interesting that the only other significant matches detected in the search were three different enzymes that can hydrolyze polysaccharides to disaccharides: Klebsiella pneumoniae cyclodextrin-glycosltransferase (19), Bacillus polymyxa β-amylase (20) and Bacillus subtilis amylase (21). These show short regions of similarity to H, D and L and to yeast maltase (data not shown) suggesting that maltases are related by descent to enzymes that yield maltose as a product.

The NH_2-terminal 10-20 amino acids of H, D and L that form the highly hydrophobic putative signal peptides extend just beyond the corresponding NH_2-terminal region of yeast maltase in the alignment, consistent with the cytoplasmic localization of yeast maltase and the prediction by Snyder and Davidson (17) that H, D and L encode secreted proteins. Furthermore, their prediction that these proteins might be involved in digestion is consistent with them being maltases, since digestion of starch requires maltase, in addition to α-amylase, to produce glucose for absorption in the gut. Our finding that the Drosophila maltases are likely to be H, D and L should aid in the understanding of carbohydrate metabolism and its regulation in an animal.

A Sarcophaga protein found in the fat body resembles Drosophila alcohol dehydrogenase

The flesh-fly, Sarcophaga pergrina, expresses an abundant mRNA in fat body encoding a 25 kD protein of unknown function (22). Our search of GenBank using this protein as probe reveals regions of highly significant similarity to the alcohol dehydrogenases (ADHs) of various Drosophila species (Figure 3). Further evidence that the 25 kD protein and the ADHs are derived from a common ancestor is that the two Adh introns in coding sequence correspond precisely in position to two of the three introns in the Sarcophaga gene encoding the 25 kD protein. However, the 25 kD protein is not likely to be an ADH. The extent of sequence similarity, 37%, is much less than one might expect if the 25 kD protein were an ADH, considering that Drosophila and Sarcophaga are both dipteran flies and that ADHs from even distant Drosophila species are nearly identical in amino acid sequence. Also, the Sarcophaga protein is about 70 amino acids shorter than Drosophila ADH at the COOH-terminal. Therefore, this



3. a) Sarcophaga 25 kD fat body protein (Sar25k) aligned with alcohol dehydrogenases from D. melanogaster (DmADH) and D. pseudoobscura (DpADH). Matches between the 25 kD protein and D. melanogaster ADH are indicated by asterisks. Differences between D. melanogaster ADH and the D. pseudoobscura proteins are shown. The ADH-like protein from D. pseudoobscura (DpADH1) is also shown, where residues that align with matching residues in the 25 kD protein are underlined.

25 kD protein is likely to have a different function, perhaps involved in physiological degradation of the fat body during metamorphosis as proposed earlier (23).

Drosophila ADH has been somewhat of a puzzle when compared to ADHs from other organisms, as its amino acid sequence does not align significantly with those of the other ADHs (24). Yet ADHs from organisms as distant as yeast and mammals are clearly conserved in primary structure. A reasonable explanation is that Drosophila ADH evolved more recently from a different protein, perhaps after loss of the "universal" Adh gene sequence. Schaeffer and Aquadro (25) have described a copy of an Adh-like gene in tandem with the Drosophila pseudoobscura Adh gene but highly diverged from it. This Adh-like gene has introns in the same position as in Adh, but it does not appear to encode an ADH. It is interesting that the Sarcophaga 25 kD protein, when aligned with the ADH-like protein, shows sequence similarities in some instances where both differ from ADH (underlined residues in Figure 3). Therefore, it is possible that the tandem D. pseudoobscura genes arose from an ancestral sequence similar to that encoding the 25 kD protein, perhaps with a similar function.

A protein encoded at the hisT operon resembles aspartate β-semialdehyde dehyrogenase

The hisT operon encodes four proteins, only two of which have identified functions (26). The protein product of hisT, pseudouridine synthase, is encoded by the third gene in the operon. The first gene in the operon

```
usg-1   MSEGVNIAVLGATGAVGEALLETLAERQFPVGEIYALARNESAGEQLRFGGKTITVQDAAEFDVTQAQLAFFVAGKEATAAVVEEATNSGCLVIDSSGLF  100
Smasd   MGYTVAIVGATGAVGTRMIQQLEQSTLPVDKVRLLSSSRSAGKVLQYKDQDVTVELTTKDSFEAVDIALFSAGGSVSAKFAPYAVKAGAVVVDNTSHF   98

usg-1   ALEPDVPLVVPEVNPFVLTDYRNRNVIAVPDSLTSQLLAALKPLIDQGGLSRISVTSLISASAQGKKAVDALAGQSAKLLN-GIPID--EEDFFGR----  193
Smasd   RQNPDVPLVVPEVNAYAMDA--HNGIIACPNCSTIQMMVALEPIRQKWGLSRVIVSTYQAVSGAGQSAINETVREIKEVVNDGVDPKAVHADIFPSGGDK  198

usg-1   ---QLAFNMLPLLP--DSEGSVREERRIVDEVRKILQDEGLMISASVVQAPVFYGHAQMVNFEALRPLAAEEARDAFVQGEDIVLSEE--NEFPTQVGDA  286
Smasd   KHYPIAFNALAQIDVFTDNDYTYEEMKMTNETKKIMEEPELPVSAHCVRVPILFSHSEAVYIETKDVAPIEEVKAAIAAFPGAVLEDDIKHQIYPQAANA  296

usg-1   SGTPHLSVGCVRNDYGMPEQVQFVSVADNVRFGGALMAVKIAEKLVQEYLY                                                    337
Smasd   VGSRTF-VGRIRKDLDIENGIHMVVVSDNLLKGAAWNSIITANRLHERGLVRSTSELKFELK                                           357
```

4. a) Alignment of E. coli usg-1 and S. mutans asd predicted proteins.

corresponds to the previously identified pdxB gene encoding an enzymatic activity in the pyridoxal phosphate pathway. Determining the function of the second gene in the operon, usg-1, is problematical, since the two flanking genes are involved in quite different pathways and since no phenotype is associated with absence of its product (27). Using the predicted usg-1 amino acid sequence (28) as probe against GenBank, we have found that it is very similar to the Streptococcus mutans asd gene product (29). An excellent alignment of the two sequences can be made, with relatively few insertions or deletions (Figure 4), suggesting that these two proteins are functionally homologous. The S. mutans gene product is thought to be aspartate β-semialdehyde dehydrogenase, an amino acid biosynthetic enzyme encoded at the asd locus in E. coli (30). However, the S. mutans asd protein is considerably more similar to E. coli usg-1 than it is to E. coli asd (28% versus 18% isology). Perhaps both the S. mutans asd and the E. coli usg-1 proteins can catalyze the aspartate β-semialdehyde dehydrogenation reaction in enzyme assays, but utilize a different substrate in vivo.

Arps and Winkler (26) have proposed that the relationship between hisT and pdxB can be understood if the operon is involved in some way in global regulation of amino acid biosynthesis, since the two pathways lead to products that are important in different ways for several different amino acid biosynthetic operons. By their reasoning, the other two genes should also be involved in similar pathways. Our evidence that one of these genes likely encodes an enzyme capable of catalyzing the asd reaction allows a test of their hypothesis and could be an important clue to a deeper understanding of metabolic regulation.

E. coli tyrosine repressor is related to nitrogen regulatory genes

The tyrR protein of E. coli regulates the expression of at least eight

```
tyrR   MRLEVFCEDRLGLTRELLDLLVLRGIDLRGIEIDPIGRIYLNFAELEFESPSSLMAEIRRIAGVTDVRTVPVWMPSEREHLALSALLEALPEPVLSVDMKS  100
ntrC                                                                                    MQRGIAWIVDDDSSIRWVLERALTGAGLSCT   31

tyrR   KVDMANPASCQLFGQKLDRLRNHTAAQLINGFNFLRVLESEPQDSHNEHVVINGQNFLMEITPVYLQDENDQHVLTGAVVMLRSTIRMGR-QLQNVAAQD  199
ntrC   TFESGNEVLDALTTKTPDVLLSDIRMPGMDGLALLK--QIKQRHPMLPVIIMTAHSDLDAAVSAYQQGAFDYLPKPFDIDEAVALVDRAISHYQEQQQPR  129

tyrR   VSAFSQIVAVSPKMKHVVEQA----QKLAMLSAPLLITGDTGTGKDLFAYACHQASPRAGKPYLALNCASIPEDAVESELFGH-------APEGKKGFFE  288
ntrC   NAPINSPTADIIGERPAMQDVFRIIGRLSRSSISVLINGESGTGKELVAHALHRHSPRAKAPFIALNMAAIPKDLIESELFGHEKGAFTGANTVRQGRFE  229

tyrR   QANGGSVLLDEIGEMSPRMQAKLLRFLNDGTFRRVGEDHEVHVDVRVICATQKNLVELVQKGMFREDLYYRLNVLTLNLPPLRDCPQDIMPLTELFVARF  388
ntrC   QADGGTLFLDEIGDMPLDVQTRLLRVLADGQFYRVGGYAPVKVDVRIIAATHQNLELRVQEGKFREDLFHRLNVIRVHLPPLRERREDIPRLARHFLQIA  329

tyrR   ADEQGVPRPKLAADLNTVLTRYAWPGNVRQLKNAIYRALTQLDGYELRPQDILLPDYDAATVAVAKMRWKVRWTKSPAVLNARY                  472
ntrC   ARELGVEAKQLHPETEMALTRLAWPGNVRQLENTCRWLTVMAAGQEVLTQDLPSELFETAIPDNPTQMLPDSWATLLGQWADRALRSGHQNLLSEAQPEM  429

ntrC   ERTLLTTALRHTQGHKQEAARLLGWGRNTLTRKLKELGME                                                            469
```

5. Alignment of E. coli tyrR and K. pneumoniae ntrC proteins.

unlinked operons involved in synthesis of the aromatic amino acids, mostly in response to the levels of phenylalanine, tyrosine and tryptophan (32). The existence of similar sequences upstream of the autogenously regulated tyrR operon and upstream of operons that are regulated by tyrR suggests that the tyrR protein acts by recognizing a sequence termed the "TYR R box" (33). Using the tyrR amino acid sequence as probe in a search of GenBank, we have found that the tyrR protein belongs to a family of proteins that regulate transcription of operons involved in nitrogen utilization and fixation. The tyrR repressor aligns with the ntrC protein, involved in general nitrogen regulation in the enteric bacterium, Klebsiella pneumoniae (Figure 5), with the K. pneumoniae nifA protein, involved in the regulation of nitrogen fixation, and with proteins involved in the regulation of nitrogen fixation in the soil bacterium, Rhizobium meliloti. The similarities between the Klebsiella and Rhizobium proteins had been demonstrated previously (34-36).

The tyrR and K. pneumoniae ntrC proteins align only within the previously identified middle domain of the latter, with 53% identical amino acids between tyrR residues 231-421 and ntrC residues 165-362. This is also the only extensive region that is conserved between ntrC and nifA (34-36), with 57% identical amino acids. The equivalent location of this conserved domain in tyrR, ntrC and nifA suggests that its involvement in transcriptional regulation is similar for the three proteins.

Other matches

In the above five cases, sequences of unknown or poorly understood function

were successfully matched with better understood proteins, leading to insights of biological significance. As well-studied protein sequences are likely to be present in protein databases, it is not surprising that three of the five proteins detected were also found in a recent protein database (Table 1). These could have been detected using a conventional searching strategy. However, in several cases, we found similarities in which the detected sequence was absent from the protein databases. These are listed in Table I and are described below.

Three examples (Table I, 6-8) reveal likely proteins that had not been identified as ORFs in earlier publications. One previously unrecognized ORF at one end of the segment that includes the E. coli tyrT locus (37) aligns with the COOH-terminal portion of GAR transformylases from different organisms. It is not known whether this is a functional locus, representing

TABLE I

Summary of previously unknown matches detected in this study

	Probe sequence	Significant Match	Isology	Aligned Residues
1.	D. melanogaster F1 protein	Elongation factor Tu[a] (several organisms)	78-89%	462
2.	D. melanogaster HDL proteins	S. carlesbergensis maltase	33%	495
3.	Sarcophaga fat body 25 kD protein	Drosophila alcohol dehydrogenase (ADH)[a]	37%	184
4.	E. coli usg-1 unknown protein	S. mutans aspartate semi-aldehyde dehydrogenase	27%	414
5.	E. coli tyrR repressor	K. pneumoniae ntrC, nifA regulatory proteins[a]	30%	413
6.	Drosophila GAR transformylase	E. coli upstream tyrT	27%	131
7.	S. cerevisiae TRP2,3 protein	H. polymorpha MOX downstream protein	66%	367
8.	E. coli thyA protein	VZV partial ORF next to glycoprotein gene	45%	90
9.	E. coli folC upstream protein	Putative chloroplast-encoded protein	39%	291
10.	E. coli pheA upstream protein	K. pneumoniae ntrA downstream fragment	37%	75

[a]Sequence present in the NBRF-PIR 14.0 protein database.

the second E. coli gene for this enzyme that had been predicted by earlier work (38). Here, the portion of the sequence encoding the expected NH_2 half of the putative protein has not been determined. Another previously unrecognized ORF downstream of the Hansenula polymorpha MOX gene (39) encodes anthranilate synthase:indole-3-glycerol phosphate synthase, as it is identical to the S. cerevisiae amino acid sequence for this enzymatic activity (40) at 66% of its predicted residues. A third example is a stretch of 273 nucleotides just downstream of a predicted glycoprotein encoded by Varicella-zoster virus (41) that encodes 108 amino acids on the other DNA strand. Detected using an E. coli thyA protein probe, this sequence is identical to the COOH-terminal 91 amino acids of human thymidylate synthase (42) at 70% of the aligned residues.

In two other cases of highly significant matches (Table I, 9-10), both predicted amino acid sequences are unknown. One is a match between an ORF upstream of the E. coli folC gene, apparently encoding the first gene of the operon (43), and an unknown ORF in both liverwort and tobacco chloroplast genomes (44, 45). The other is a match between a 113 amino acid unknown ORF upstream of the E. coli pheA operon (46) and an unidentified ORF downstream of the Klebsiella pneumoniae ntrA gene (47). The Klebsiella sequence entry ends after alignment of 28 of the 75 amino acids forming the likely NH_2-terminal. This example demonstrates that partial amino acid sequences, too short to be recognized as likely protein-coding regions, are sufficient to be detected by this approach.

DISCUSSION

We have shown that simple searching of a nucleotide sequence database using amino acid sequences as probes is capable of detecting significant homologies. Most of the examples that we used to illustrate this approach were cases in which sequences of unknown function were successfully matched with known genes, leading to identification of the unknown.

Our choice of probes was neither systematic nor random, so that it is difficult to extrapolate our experience to other unknown sequences. However, there are several hundred recognized but unknown ORFs compiled in the DNA databases. Only a small fraction of these have been examined thus far. In addition, a large number of unrecognized protein-coding sequences, such as the three such examples we described, probably also exist in the data; these could be detected and perhaps identified using our strategy. This possibility is particularly applicable to the genomes of higher eukaryotes, where ORFs are

usually fragmented by introns, and where introns themselves are sometimes sites for entire genes (48-50).

It is worth noting that sequencing errors, even those that cause frameshifts, are not likely to obliterate an amino acid sequence similarity when this method is used, since all frames are examined and since stop codons are ignored. Such frameshifting can occur in databases; for example, in the H. polymorpha MOX unrecognized ORF, conceptual insertion of a single base after nucleotide 2900 shifts the frame to continue alignment for another 117 amino acids, whereupon the published sequence ends (39). Our method requires only a short region of fairly accurate sequence, as illustrated by the homeobox example described above.

The ease with which we were able to find biologically important matches for published sequences leads us to believe that this approach has not been widely used. It appears that investigators who generate new sequence data generally limit their searches to protein databases, which are necessarily far less complete than the DNA databases. For example, some of the probe sequences that were used in our study are absent from the most recent version of SWISS-PROT 6 (January, 1988), a composite database that includes proteins present in the NBRF-PIR 14.0 database as well as those deduced from sequences present in the EMBL 13.0 DNA database (51). Among the missing proteins are Drosophila F1 and HDL and Sarcophaga fat body proteins, all of which have been present in the DNA databases for years. Since most protein database sequences are now extracted from the DNA databases, the translated version necessarily appears later than the DNA entry. For example, the E. coli TyrR protein sequence, obtained from a DNA database, was not present in the NBRF-PIR protein database when we carried out our search. However, this sequence has since been added and the similarity to nitrogen fixation proteins has been noted.

The DNA database searching approach is not limited to the particular searching protocol that we used in this study. For example, a large diverged family of bacterial regulatory proteins has been detected using a window of 90 rather than 30 (52). In this study, a consensus sequence was derived and used as probe in database searches to further improve sensitivity. Another modification of the approach is to translate all six possible reading frames of a DNA segment and use the composite to search entire DNA databases. For example, when this was done with the entire transposable element TC1 from C. elegans (53), a previously unknown relationship between a TC1 ORF and one within a D. melanogaster HB transposable element (54) was detected (S. H. and R. Plasterk, submitted for publication).

In summary, we have shown that simple searching of nucleotide sequence databases is an effective method for detecting correspondences between proteins. This method should become increasingly valuable to individual researchers as such databases expand.

ACKNOWLEDGEMENTS

We thank F. Ausubel, N. Davidson, J. Levy, J. Marmur, S. Potter, W. Ruyechan, S. Schaeffer, B. Shane, J. Smith, M. Snyder, A. Wahba and M. Winkler for advising us that relationships reported here were previously unknown to them. S. H. received support from NIH grant GM-29009. We thank Ed Lang and Connie Furlong for contributions of hardware and software.

*To whom correspondence should be addressed

REFERENCES

1. Doolittle, R. F., Hunkapiller, M. W., Hood, L. E., Devare, S. G., Robbins, K. C., Aaronson, S. A. and Antoniades, H. N. (1983) Science 221, 275-276.
2. Vogt, P. K., Bos, T. J. and Doolittle, R. F. (1987) Proc. Natl. Acad. Sci. USA 84, 3316-3319.
3. Claverie, J. M. and Bricault, L. (1986) Proteins 1, 60-65.
4. Dayhoff, M. (1978) Atlas of protein sequence and structure, vol. 5, suppl. 3, National Biomedical Research, Washington, D.C., pp.345-358.
5. Dayhoff, M. O., Barker, W. C. and Hunt, L. T. (1983) Meth. Enzymol. 91, 524-545.
6. Wilbur, W. J. and Lipman, D. J. (1983) Proc. Natl. Acad. Sci. USA 80, 726-730.
7. Lipman, D. J. and Pearson, W. R. (1985) Science 227, 1435-1441.
8. McGinnis, W., Garber, R. L., Wirz, J., Kuroiwa, A. and Gehring, W. J. (1984) Cell 37, 403-408.
9. Scott, M. P. and Weiner, A. J. (1984) Proc Nat Acad Sci USA 81, 4115-4119.
10. Walldorf, U., Hovemann, B. and Bautz, E. K. F. (1985) Proc. Natl. Acad. Sci. USA 82, 5795-5799.
11. Lenstra, J. A., Vliet, A. V., Arnberg, A. C., Hemert, F. J. V. and Moller, W. (1986) Eur. J. Biochem. 155, 475-483.
12. Brands, J. H. G. M., Maassen, A., Van Hemert, F. J., Amons, R. and Moller, W. (1986) Eur. J. Biochem. 155, 167-171.
13. Nagata, S., Nagashima, K., Tsunetsugu-Yokota, Y., Fujimura, K., Miyazaki, M. and Kaziro, Y. (1984) EMBO J. 3, 1825-1830.
14. Cottrelle, P., Thiele, D., Price, V. L., Memet, S., Micouin, J.-Y., Marck, C., Buhler, J.-M., Sentenac, A. and Fromageot, P. (1985) J. Biol. Chem. 260, 3090-3096.
15. Linz, J. E., Lira, L. M. and Sypherd, P. S. (1986) J. Biol. Chem. 261, 15022-15029.
16. Roth, W. W., Bragg, P. W., Corrias, M. V., Reddy, N. S., Dholakia, J. N. and Wahba, A. J. (1987) Mol. and Cell. Biol. 7, 3929-3936.
17. Snyder, M. and Davidson, N. (1983) J. Mol. Biol. 166, 101-118.
18. Hong, S. H. and Marmur, J. (1986) Gene 41, 75-84.
19. Binder, F., Huber, O. and Boeck, A. (1986) Gene 47, 269-277.
20. Kawazu, T. Nakanishi, Y., Uozumi, N., Sasaki, T., Yamagata, H., Tsukagoshi, N. and Udaka, S. (1987) J. Bacteriol. 169, 1564-1570.

21. Yang, M., Gallizzi, A. and Henner, D. (1983) Nucl. Acids Res. 11, 237-
22. Tamura, H,. Tahara, T., Kuroiwa, A., Obinata, M. and Natori, S. (1983) Dev. Biol. 99, 145-151
23. Matsumoto, N., Sekimizu, D., Soma, G.-I., Ohmura, Y., Andoh, T., Nakanishi, Y., Obinata, M. and Natori, S. (1985) J. Biochem. 97, 1501-1508.
24. Thatcher, D.R. (1980) Biochem. J. 187, 875-886.
25. Schaeffer, S. W. and Aquadro, C. F. (1987) Genetics 117, 61-73.
26. Arps, P. J. and Winkler, M. E. (1987) J. Bacteriol. 169, 1071-1079.
27. Arps, P. J. and Winkler, M. E. (1986) J. Bacteriol. 169, 1061-1070.
28. Arps, P. J., Marvel, C. C., Rubin, B. C., Tolan, D. R., Penhoet, E. E. and Winkler, M. E. (1985) Nucl. Acids Res. 13, 5297-5315.
29. Cardineau, G. A. and Curtiss, R. III (1987) J. Biol. Chem. 262, 3344-3353.
30. Haziza, C. Stragier, P. and Patte, J.-C. (1982) EMBO J. 1, 379-384.
31. Jagusztyn-Krynicka, KE. K., Smorawinska, M. and Curtiss, R. III (1982) J. Gen. Microbiol. 128, 1135-1145.
32. Cornish, E. C., Argyropoulos, Pittard, J. and Davidson, B. E. (1986) J. Biol. Chem. 261, 403-410.
33. Chye, M.-L. and Pittard, J. (1987) J. Bacteriol. 169, 386-393.
34. Buikema, W., J., Szeto, W. W., Lemley, P. V., Orme-Johnson, W. H., and Ausubel, F. M. (1985) Nucl. Acids Res. 13, 4539-4555.
35. Drummond, M., Whitty, P. and Wootton, J. (1986) EMBO J. 5, 441-447.
36. Szeto, W. W., Nixon, B. T., Ronson, C. W. and Ausubel, F. M. (1987) J. Bacteriol. 169, 1423-1432.
37. Sprinzl, M. and Gauss, D. H. (1983) Nucl. Acids Res. 11, r55-r103.
38. Smith, J. M. and Daum, H. A. III (1987) J. Biol. Chem. 262, 10565-10569.
39. Ledeboer, A. M., Edens, L., Maat, J., Visser, C., Bos, J. W., Verrips, C. T., Janowicz, Z., Eckart, M., Roggenkamp, R. and Hollenberg, C. P. (1985) Nucl. Acids Res. 13, 3063-3082.
40. Zalkin, H., Paluh, J. L., van Cleemput, M., Moyes, W. S. and Yanofsky, C. (1984) J. Biol. Chem. 259, 3985-3992.
41. Kinchington, P. R., Reneick, J, Ostrove, J. M., Straus, S. E., Ruyechan, W. T. and Hay, J. (1986) J. Virology 59, 660-668.
42. Takeishi, K., Kaneda, S., Ayusawa, D., Shimizu, K., Gotoh, O. and Seno, T. (1985) Nucl. Acids Res. 13, 2035-2043.
43. Bognar, A. L., Osborne, C. and Shane, B. (1987) J. Biol. Chem. 262, 12337-12343.
44. Shinozaki, K., Ohme, M., Tanaka, M., Waksugi, T., Hayashida, N., Matsubayashi, T., Zaita, N., Chunwongse, J., Obokata, J., Yamaguchi-Shinozaki, K., Ohto, C., Toazawa, K., Meng, B. Y., Sugita, M., Deno, H., Kamogashira, T., Yamada, K., Kusuda, J., Takaiwa, F., Dato, A., Tohdoh, N., Shimada, H. and Sugiura, M. (1986) EMBO J. 5, 2043-2049.
45. Ohyama, K., Fukuzawa, H., Kohchi, T., Shirai, H., Sano, T., Sano, S., Umesono, K., Shiki, Y., Takeuchi, M., Chang, Z., Aota, S.-I., Inokuchi, H. and Ozeki, H. (1986) Nature 322, 572-574.
46. Hudson, G. S. and Davidson, B. E. (1984) J. Mol. Biol. 180, 1023-1051.
47. Merrick, M. J. and Gibbins, J. R. (1985) Nucl. Acids Res. 13, 7607-7620.
48. Henikoff, S., Keene, M. A., Fechtel, K. and Fristrom, J. W. (1986) Cell 44, 33-42.
49. Adelman, J. P., Bond, C. T., Douglass, J. and Herbert, E. (1987) Science 235, 1514-1517.
50. Chen, C., Malone, T., Beckendorf, S. K. and Davis, R. L. (1987) Nature 329, 721-724.
51. Bairoch, A., Cameron, G., Hazledine, D, Jones, S., Stoesser, G., Kahn, P. Apweiler, R. and Breun, G. (1988) SWISS-PROT Protein Sequence Database, Release 6.
52. Henikoff, S., Haughn, G. W., Calvo, J. M. and Wallace, J. C. (1988) Proc. Natl. Acad. Sci. USA, In press.
53. Herman, R. K. and Shaw, J. E. (1987) Trends in Genetics 3, 222-225.
54. Brierley, H. L. and Potter, S. S. (1985) Nucl. Acids Res. 13, 485-500.