Proc. Natl. Acad. Sci. USA
Vol. 79, pp. 2836–2839, May 1982
Biochemistry

Viral *src* gene products are related to the catalytic chain of mammalian cAMP-dependent protein kinase

(Rous sarcoma virus/Moloney murine sarcoma virus/transforming protein/protein homologies)

W. C. BARKER AND M. O. DAYHOFF

National Biomedical Research Foundation, Georgetown University Medical Center, Washington D.C. 20007

Communicated by Edmond H. Fischer, February 1, 1982

ABSTRACT **The transforming protein sequences translated from the Rous avian and Moloney murine sarcoma virus *src* genes are shown to be related to the catalytic chain of bovine cAMP-dependent protein kinase (ATP:protein phosphotransferase, EC 2.7.1.37). The avian transforming protein, also a protein kinase, shows greatest homology with the bovine protein kinase in the carboxyl-terminal half, where the protein kinase activity is localized. Moreover, lysine occurs in the inferred transforming protein sequences at the position homologous with the proposed ATP-binding lysine of the bovine protein kinase. This relationship is consistent with the hypothesis that the *src* genes originated in the host genomes, in which they are members of a superfamily of distantly related protein kinases that are normal constituents of mammalian cells. In the host, these sequences are much more highly conserved than in the viruses.**

The complete sequence of the catalytic chain of cAMP-dependent protein kinase (ATP:protein phosphotransferase, EC 2.7.1.37) from bovine cardiac muscle has recently been determined (1). This enzyme plays a central role in many cAMP-dependent mechanisms in eukaryotic organisms. The inactive form of this enzyme is a tetramer consisting of two catalytic chains and two regulatory chains. cAMP binds to the regulatory chain, promoting dissociation of the molecule. The catalytic chains are then free to catalyze the phosphorylation of susceptible serine residues in protein substrates. Each catalytic chain consists of 349 residues, including one phosphorylated threonine (residue 196) and one phosphorylated serine (residue 337). Lysine-71 is proposed to be the ATP binding site (1) on the basis of homology with the ATP-binding region of the catalytic chain of porcine protein kinase.

In the process of routinely screening new sequences having no known relatedness to previously determined protein sequences, we noted regions of similarity of the catalytic chain of bovine cAMP-dependent protein kinase (BOV-PK) to the inferred amino acid sequences of the *src* gene products (transforming proteins) of Moloney murine sarcoma virus (MMSV) and of Rous avian sarcoma virus, Schmidt–Ruppin strain (RSV-SR). The transforming protein sequences were translated from viral nucleic acid sequences (2–4) and the homology of the murine virus sequence with the carboxyl-terminal 65% of the avian virus sequence has been noted (4). Moreover, the avian virus transforming protein has been shown to have protein kinase activity (5), catalyzing the phosphorylation of tyrosine residues in target proteins (6). We have tested the cAMP-dependent protein kinase and the *src* transforming protein amino acid sequences for relatedness to one another and find that they are indeed distantly related.

After these studies were in progress, the complete nucleotide sequence of the genome of Rous sarcoma virus, Prague C strain (RSV-PC), became available (Dennis Schwartz, personal communication). Although the RSV-PC *src* gene sequence is ≈95% identical with the other avian *src* gene sequence, there are some insertions and deletions that affect the reading frame in the protein translation. The translated sequences are very different in the regions corresponding to residues 20–189, 222–227, 268–308, 370–376, and 507–514 of the RSV-SR *src* gene translation (3). The remaining regions of these sequences are ≈97% identical. Both sequences were used in the studies reported here. In addition, the two reported versions of the MMSV *src* gene sequence (2, 4) produce very different amino acid translations after residue 349. The longer sequence (2) was (arbitrarily) used in the studies reported here.

COMPUTER METHODS

The program SEARCH (7) compares a test sequence of, for example, 25 residues with all 25-residue segments of the protein sequences in our data base and with the progressively shorter segments at the ends of those sequences. The distribution of scores from unrelated sequences is approximately normal; related segments appear in an abnormally long tail of high scores. Typically, for a 25-residue segment, all corresponding sequences of the same protein family (sequences <50% different) appear above the distribution of scores of unrelated segments. Usually, a number of the more distantly related sequences are also above this distribution, whereas the rest are within the upper tail of scores from unrelated segments. Program SEARCH is used as a screening procedure to identify possible candidates for relationship to the test sequence. Unless the homology is obvious and extensive, we use another computer program to assess the probability that the similarities observed have occurred by chance.

The program RELATE (7) is suitable for testing proteins of differing lengths for relatedness. It compares all possible segments of a given length (25 residues in the studies reported here) from one sequence with all segments of the same length from a second sequence (8). A segment score is accumulated from the pair scores of the amino acids occupying corresponding positions within the two segments. These pair scores are specified by a scoring matrix; we use the mutation data matrix (9). A numerical property of the distribution of segment scores is determined for the given sequences and for a large number (usually 100) of comparisons of random permutations of the sequences. We use for the numerical property the mean of a num-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

Abbreviations: BOV-PK, catalytic chain of bovine cAMP-dependent protein kinase; MMSV, Moloney murine sarcoma virus; RSV-SR, Rous sarcoma virus, Schmidt–Ruppin strain; RSV-PC, Rous sarcoma virus, Prague C strain.

ber of the highest segment scores. The number of segment scores used is equal to the number of scores to be expected from corresponding segments if the sequences are related (the length of the shorter sequence minus one less than the segment length). The segment comparison score is calculated as the difference between this numerical property determined from comparison of the real sequences and the average value determined from the many pairs of permuted sequences divided by the SD of the values from the randomized sequences. The segment comparison score is thus expressed in SD units, and the probability of occurrence of a particular score by chance can be found. We generally assume that a score >5 SD indicates evolutionary relatedness of two proteins and scores between 3 and 5 SD support relationship if there are other indications such as similarity of function.

RESULTS

The sequence of the catalytic chain of bovine cAMP-dependent protein kinase was entered into the protein data base of the *Atlas of Protein Sequence and Structure* in April 1981. Because it was not known to be related to any other proteins whose sequences have been determined, we used the SEARCH program to compare consecutive segments (residues 1–25, 26–50, etc.) against the current protein data base (containing 1,560 sequences) using the mutation data scoring matrix. The BOV-PK segment consisting of residues 151–175 matched two segments in the data base with unusually high scores, residues 375–399 of the inferred amino acid sequence of the RSV-SR transforming protein and residues 215–239 of the inferred amino acid sequence of the MMSV transforming protein. For two other segments of the protein kinase, the avian virus protein scored among the top 25 scores of 192,000 segments compared. The transforming proteins were tested further for possible relatedness to the protein kinase by using the program RELATE.

The segment comparison scores from intercomparisons of BOV-PK and the MMSV, RSV-SR, and RSV-PC transforming proteins are shown in Table 1. The probabilities of BOV-PK being so similar by chance to the MMSV and RSV transforming proteins are $<10^{-3}$ and $<10^{-10}$, respectively. The probability of the mouse and avian transforming proteins being so similar by chance is $<10^{-9}$. On the basis of these results, we have assigned these proteins to the same superfamily.

Again using the program RELATE, we tested the protein kinase and the viral transforming proteins against a number of other amino acid sequences inferred from viral nucleotide sequences, including the middle-size tumor antigen of polyoma virus, which is also reported to have protein kinase activity (10). None of the sequences tested had significant scores against all (or even two) of the proteins in this superfamily. The bovine protein kinase was also tested against the seven other kinases in the data base: pig adenylate kinase, horse phosphoglycerate kinase, herpes simplex thymidine kinase, *Escherichia coli* aspartokinase I/homoserine dehydrogenase I and homoserine kinase, *Bacillus stearothermophilus* phosphofructokinase, and bacteriophage T7 protein kinase. The scores ranged from −0.3 SD to 2.6 SD.

Table 1. Segment comparison scores of BOV-PK and viral transforming proteins (*src* gene products)

	SD units	
	BOV-PK	MMSV
BOV-PK	—	3.1
MMSV transforming protein	3.1	—
RSV-SR transforming protein	7.5	6.3
RSV-PC transforming protein	7.8	7.0

Part of the output of the RELATE program, a list of the highest scoring segment comparisons, serves as a guide to locating the regions of greatest sequence similarity, which for BOV-PK and both of the transforming proteins corresponds to positions 129–174 of the alignment in Fig. 1. In the alignment, which includes >60% of the sequence of BOV-PK, there are 21–24% identities between any two of the sequences. Outside of the regions shown in Fig. 1, the homology is weak and few identities occurred in any attempted alignment of the three sequences. The BOV-PK sequence and the MMSV transforming protein sequence are distantly related to the carboxyl-terminal portion of the larger RSV-PC sequence, which is the enzymatically active region of the molecule (11). The inferred RSV-SR sequence is nearly identical with the inferred RSV-PC sequence except for alignment positions 8–51 and 130–136. In these two regions, the RSV-PC sequence is more similar to the BOV-PK and inferred MMSV transforming protein sequences, so we have used it in the figure.

The regions shown in the alignment include the lysine at position 39, which is the proposed ATP-binding site of BOV-PK (1), residue 71 in the protein chain; lysine also occurs in the other sequences at the homologous position. A short distance before the conserved lysine is a region in which six of nine consecutive residues (positions 16–24) are conserved in all of these sequences. This pattern, Leu-Gly-X-Gly-X-Phe-Gly-X-Val, does not occur in any other sequence in the Protein Sequence Data Base. The phosphorylated threonine of BOV-PK (1), residue 196 in the protein chain, occurs at position 180 of the alignment. Remarkably, in the avian virus transforming protein, it is the adjacent tyrosine at position 179 of the alignment that is phosphorylated (12), although threonine is found at position 180.

DISCUSSION

One criticism of these results that may have occurred to the reader is that amino acid sequences inferred from nucleic acid sequences may contain gross errors resulting from rather minor errors in the nucleic acid sequence. For instance, if one or two nucleotides are missing from the sequence, the amino acid sequence will be translated in the wrong reading frame until perhaps a subsequent error restores the correct reading frame. The inferred amino acid sequences of RSV-PC and RSV-SR transforming proteins are 40% different, mainly because of this type of discrepancy in the reported nucleotide sequences. Nevertheless, the segment comparison scores of these two sequences against the BOV-PK were very similar.

The segment comparison scores are based on only a small fraction of the segment scores. For example, in comparing BOV-PK with the RSV-PC sequence, only the highest 325 of the 163,150 segment scores are used in the final calculation. Most scores from segments for which the RSV-PC and RSV-SR sequences disagree fall within the range of scores from nonhomologous segments and, therefore, they have little influence on the final score. When the discrepancies between the translated transforming protein sequences have been resolved and the sequences have been corroborated by isolation and sequence analysis of the proteins, their homology with BOV-PK may be shown to be greater than is now evident but is unlikely to be less.

The genomes of both chickens and mice contain one or two copies of a gene closely related to the viral *src* gene (13). This cellular gene (called *sarc*) is not detectably expressed during viral or otherwise induced transformation and it is evolving at a rate similar to other cellular genes (hemoglobin, for example). In contrast, genes of the endogenous leukemia viruses, which are closely related to sarcoma viruses but lack the *src* gene, are

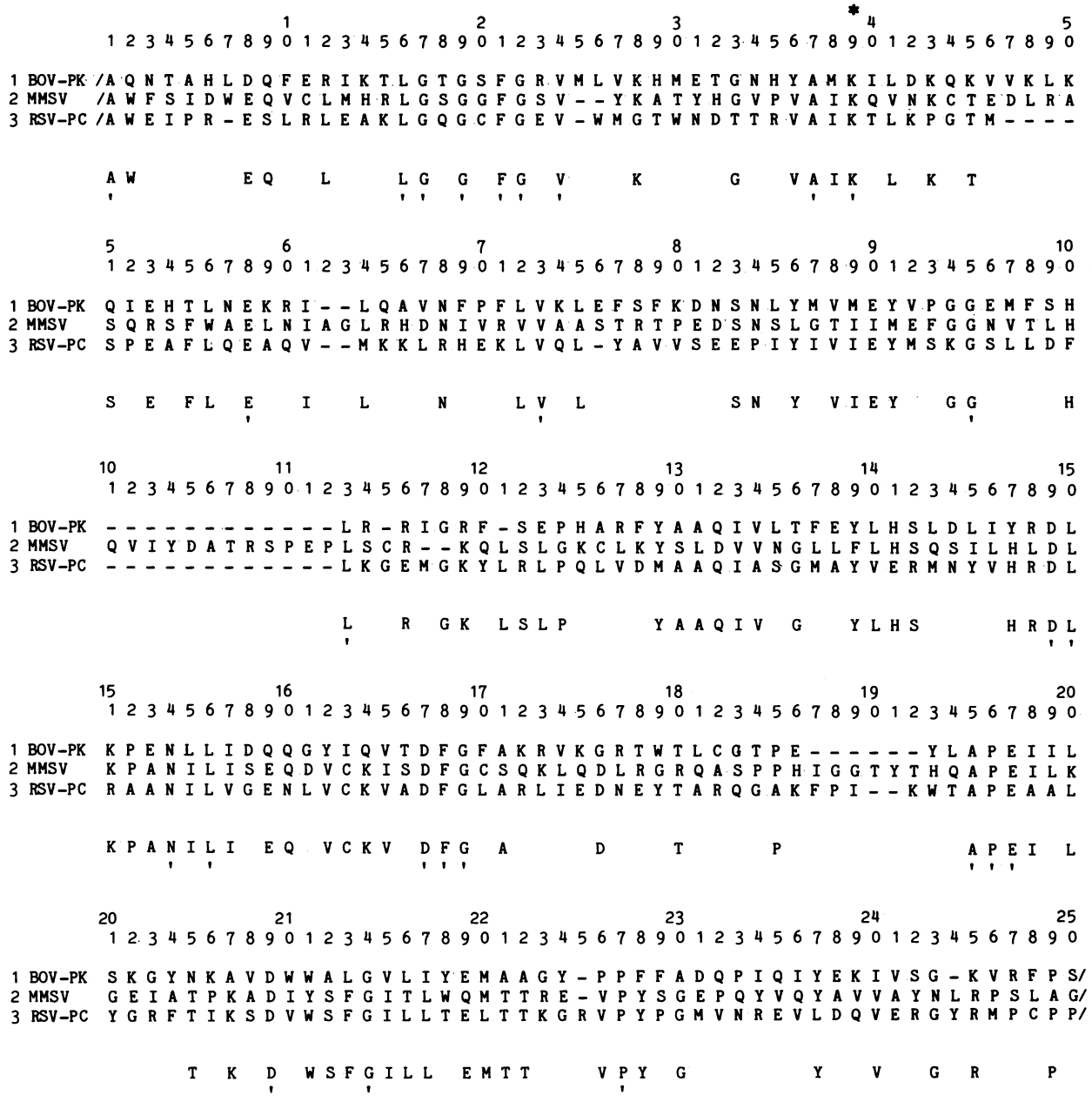

FIG. 1. Alignment of residues 33–258 of BOV-PK, residues 85–329 of the transforming protein of MMSV, and residues 259–485 of the transforming protein of RSV-PC. The lysine residue that binds ATP in BOV-PK is at position 39 in the alignment. Amino acids shown below the alignment are common to at least two of the sequences; those common to all are marked. The amino acid sequence of BOV-PK was determined directly (1); those of the transforming proteins are tentative translations of viral nucleotide sequences (ref. 2; Dennis Schwartz, personal communication). A, alanine; C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine.

present in 5–10 copies per genome and have changed considerably in even closely related species (13). Thus, it seems that the *sarc* gene is a normal cellular gene.

Several lines of evidence indicate that the *sarc* gene is functional. RNA transcripts of the *sarc* gene are found in normal avian cells in small quantities. Moreover, proteins highly related to the RSV transforming protein and having associated protein kinase activity are found in cells of other birds, rats, and humans (14). The normal function of tyrosine-specific protein kinases in mammalian cells is unknown. However, epidermal growth factor has been shown to stimulate the activity of a membrane-associated tyrosine-specific protein kinase, suggesting that such enzymes may play very specific roles in normal growth and development (15). The genes coding for these enzymes become oncogenic when their normal control of expression is deranged by incorporation into a viral genome.

One paradox is evident: the retroviruses must have developed the ability to incorporate particular host genes long ago and yet the genes in the viruses and in their present-day hosts are similar. A mechanism for the coevolution of corresponding viral and host genes is easily imagined. If the *src* gene sequence were to change much, it probably would no longer be oncogenic. However, a virus carrying such a defective *src* gene might exchange genetic information with the host, reincorporating the cellular *sarc* gene or a portion of it. Indeed, it has been shown that tumors that sometimes appear after infection with transformation-defective RSV-SR, which lacks most of the *src* gene, contain transformation-competent viruses. These re-

covered viruses contain a *src* gene that is a recombinant of the viral *src* gene and the cellular *sarc* gene (16).

We thank Dennis Schwartz and his colleagues Richard Tizard and Walter Gilbert for making their nucleotide sequence of Rous sarcoma virus, strain Prague C, available to us prior to publication. This work was supported by National Institutes of Health Grant HD09547.

1. Shoji, S., Parmelee, D. C., Wade, R. D., Kumar, S., Ericsson, L. H., Walsh, K. A., Neurath, H., Long, G. L., Demaille, J. G., Fischer, E. H. & Titani, K. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 848–851.
2. Reddy, E. P., Smith, M. J., Canaani, E., Robbins, K. C., Tronick, S. R., Zain, S. & Aaronson, S. A. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 5234–5238.
3. Czernilofsky, A. P., Levinson, A. D., Varmus, H. E., Bishop, J. M., Tischer, E. & Goodman, H. M. (1980) *Nature (London)* **287**, 198–203.
4. Van Beveren, C., Galleshaw, J. A., Jonas, V., Berns, A. J. M., Doolittle, R. F., Donoghue, D. J. & Verma, I. M. (1981) *Nature (London)* **289**, 258–262.
5. Collett, M. S. & Erikson, R. L. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 2021–2024.
6. Hunter, T. & Sefton, B. M. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 1311–1315.
7. Dayhoff, M. O. (1979) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. O. (National Biomedical Research Foundation, Washington, DC), Vol. 5, Suppl. 3, pp. 1–8.
8. Fitch, W. M (1966) *J. Mol. Biol.* **16**, 9–16.
9. Schwartz, R. M. & Dayhoff, M. O. (1979) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. O. (National Biomedical Research Foundation, Washington, DC), Vol. 5, Suppl. 3, pp. 353–358.
10. Smith, A. E., Smith, R., Griffin, B. & Fried, M. (1979) *Cell* **18**, 915–924.
11. Levinson, A. D., Courtneidge, S. A. & Bishop, J. M. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 1624–1628.
12. Neil, J. C., Ghysdael, J., Vogt, P. K. & Smart, J. E. (1981) *Nature (London)* **291**, 675–677.
13. Fischinger, P. J. (1980) in *Molecular Biology of RNA Tumor Viruses*, ed. Stephenson, J. R. (Academic, New York), pp. 163–198.
14. Oppermann, H., Levinson, A. D., Varmus, H. E., Levintow, L. & Bishop, J. M. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 1804–1808.
15. Ushiro, H. & Cohen, S. (1980) *J. Biol. Chem.* **255**, 8363–8365.
16. Karess, R. E. & Hanafusa, H. (1981) *Cell* **24**, 155–164.