ELSEVIER

FEBS Letters 343 (1994) 109–114

FEBS 13934

FEBS LETTERS

A two-domain model for the R domain of the cystic fibrosis transmembrane conductance regulator based on sequence similarities

Ann M. Dulhanty*, John R. Riordan

Cystic Fibrosis Research Development Program and Biochemistry Department, Hospital for Sick Children, 555 University Ave., Toronto, Ont., M5G 1X8, Canada

Received 10 March 1994

Abstract

CFTR belongs to a group of proteins sharing the structural motif of six transmembrane helices and a nucleotide binding domain. Unique to CFTR is the R domain, a charged cytoplasmic domain. Comparison of R domain sequences from ten species revealed that the N-terminal third is highly conserved, while the C-terminal two-thirds is poorly conserved. The R domain shows no strong sequence similarity to known proteins; however, 14 viral pol proteins show limited similarity to fragments of the R domain. Analysis revealed a relationship between the N- and C-terminal fragments of the R domain and two discontinuous fragments of the pol protein. These observations support a two-domain model for the R domain.

Key words: Alignment; CFTR; Retrovirus; Homology; Pol protein

1. Introduction

It is common to be faced with the predicted sequence of a gene product with only the slimmest clues as to its function. To obtain guidance for investigation, information may be obtained from the amino acid sequence. Alignment of the sequence from different species may indicate catalytically significant residues, and areas where structure is rigidly constrained. Searches of protein data bases may provide functional identification of the protein, if another protein of sufficient sequence similarity has been characterized. Currently, functional identification, based on sequence similarity, is possible for only one in six newly identified amino acid sequences [1]. In the absence of functional identification of the gene product, functional domains may be identified or structural information may be obtained on the basis of sequence similarity [2].

The cloning of the cystic fibrosis gene, which encodes the cystic fibrosis transmembrane conductance regulator (CFTR) protein, was accomplished by genetic methods [3]. Although involvement in ion transport was indicated [4,5], the function of CFTR was unknown when the amino acid sequence was determined. Searches through protein databases indicated that CFTR contains two nucleotide binding folds (NBF's) [6] and twelve transmembranes [7] and is related to a large super family of transporters containing these structural motifs [8]. Within this group, CFTR alone contains a highly charged, cytoplasmic domain with little sequence similarity to other known proteins. This unique, regulatory or R domain of CFTR, is phosphorylated by PKA [9,10] and it has now been demonstrated that CFTR function involves a PKA regulatable chloride conductance [11,12]. The R domain appears to have a role in gating the chloride channel which mediates this conductance [13,14].

We have now analyzed the alignment of the R domain amino acid sequences from ten animal species. This analysis and a distant relationship between the R domain and viral pol proteins which was discovered, suggests that the R domain is made up of two distinct parts.

2. Methods, results and discussion

2.1. R domain sequence conservation

Alignment of the R domain amino acid sequence from ten animal species was performed using the SEQSEE program [15]. The scoring matrix from this package emphasizes amino acid substitutions that preserve secondary structure. The R domain sequences of human [6], mouse [16], rat [17], dogfish [18], frog [19], and cow [20] consist of the amino acids encoded by exon 13 of CFTR (amino acids 591–830), while the rat, ginea pig, sheep and monkey [20] consist of amino acids 604 to 777. The

* Corresponding author. Fax: (1) (416) 813 7099.

Abbreviations: CFTR, cystic fibrosis transmembrane conductance regulator protein; PKA, cAMP dependent protein kinase; PKC, protein kinase C; NBF, nucleotide binding fold; PIR, Protein Resource Information; MMLV, Moloney murine leukemia virus.

SSDI 0014-5793(94)00319-Q

residue numbering scheme of human CFTR is used here [6].

To provide a visual representation of sequence conservation, a plot was prepared of the sequence similarity score as a function of the position in the alignment. The positions in the alignment were numbered from 1 to 251. Position numbers exceed the residue numbers in the human R domain (241), because of insertions in the shark domain and alignment gaps. To calculate the sequence similarity score for each position, the residue from each species at that position was compared with the aligned human residue. An individual score for the comparison was assigned with the SEQSEE scoring matrix [15]. This matrix gives scores for each of the possible 210 amino acid substitutions or conservations: values of 10 to 16 reflect conservation, values of 3 to 9 conservative substitution, and values of 0 to 2 non-conservative substitution. Gaps in the alignment were scored as 0. The sequence similarity score for each residue position is the sum of the individual scores for comparison of each species with the human. Where sequences from only six species were compared, the score was multiplied by 1.8 for normalization. Fig. 1A shows the results of this calculation. Positions of low sequence similarity are likely to represent turns or external loops in the R domain. It is clear that approximately the first 80 amino acids are highly conserved, showing 64% identity. The next 140 amino acids are less well conserved with 23% identity. The final 20 residues in the alignment are 68% identical. Gaps in the alignment are notable at positions 83 and 225.

2.2. *Two domain model of the R domain*

The R domain was originally defined as exon 13 of CFTR [6]. However, the high sequence similarity at the boundaries of exon 13 suggests protein domain boundaries may not coincide. To postulate domain boundaries, previously recognized domains of CFTR were included. The first NBF, as defined by alignment with other known NBF's, ends three residues prior to the beginning of the exon 13 coding sequence. Twenty-nine amino acids are predicted between the end of exon 13 coding sequence and the seventh transmembrane domain. The sequence similarity pattern of the R domain suggests two discrete domains. Differences in the degree of conservation in separate functional domains of a single polypeptide chain has been observed in other proteins [21]. We postulate, RD1, the first domain within the R domain to extend from amino acids 587 to 672 of CFTR. High homology in this domain suggests that the structural constraints are rigid. RD2, the C-terminal domain of the R domain, with less rigid structural requirements, extends from amino acids 679 to 798 of CFTR. The gap in the sequence alignment at position 225 and the abrupt change in conservation suggest that amino acid 806 of CFTR is the beginning of the RTM region, which links the R domain and the seventh transmembrane domain. Fig. 1B shows an R domain alignment with all domain boundaries highlighted.

2.3. *Conservation of phosphorylation sites*

Phosphorylation of CFTR by PKA mediates protein function. Phosphorylation by PKC also has a role in controlling protein function, but it is subtle and less clearly understood [22,23]. Most of the predicted PKA and PKC phosphorylation sites in the R domain fall in RD2 where the primary sequence is not well conserved, as shown in Fig. 1A. However, 8 of the PKA sites are highly conserved: amino acids 660, 686, 700, 712, 737, 768, 795 and 813 in CFTR. The site at residue 196 (amino acid 788) is found only in human. Similarly, most of the PKC sites are well conserved, with the exception of the site at amino acid 791. Although the C-terminal domain of the R domain is itself not well conserved, many of the phosphorylation sites are, suggesting that while the structural requirements on the domain are not stringent, phosphorylation of the domain is important for CFTR function.

2.4. *Low sequence similarity between the R domain and pol proteins*

At the time of the cloning of CFTR, no other proteins or protein domains were identified that were highly similar to the R domain [6]. This finding still holds. However, it is stricking that a low level of similarity exists between various segments of the R domain and 14 different viral pol (polymerase) proteins (Fig. 2), as revealed in searches of the Protein Resource Information (PIR) database (National Biomedical Resource Foundation, Washington, DC). The FASTA algorithm [24] identified seven retroviral pol proteins showing sequence similarity to the R domain. Seven pol proteins from a heterogeneous group of RNA genome viruses also show limited sequence similarity to the R domain. These pol proteins have been described as RNA polymerase activities, and the reverse transcriptase domain of the retroviral pol protein was originally defined by sequence similarity to the *E. coli* RNA polymerase α subunit [39].

2.5. *Significance of alignment between R domain and pol proteins*

The similarity between the R domain and pol proteins falls into a statistically wooly area. The alignment of the R domain with each of the 14 viral proteins scored between 3.3 and 7.8 standard deviations from the mean. Scores between three and six standard deviations above the mean are considered to be equally consistent with either two distantly related sequences or two sequences which are unrelated and align by chance [40]. Monte Carlo analysis was performed to address the possibility of common bias in amino acid usage but proved inconclusive [41]. Sander and Schneider have demonstrated

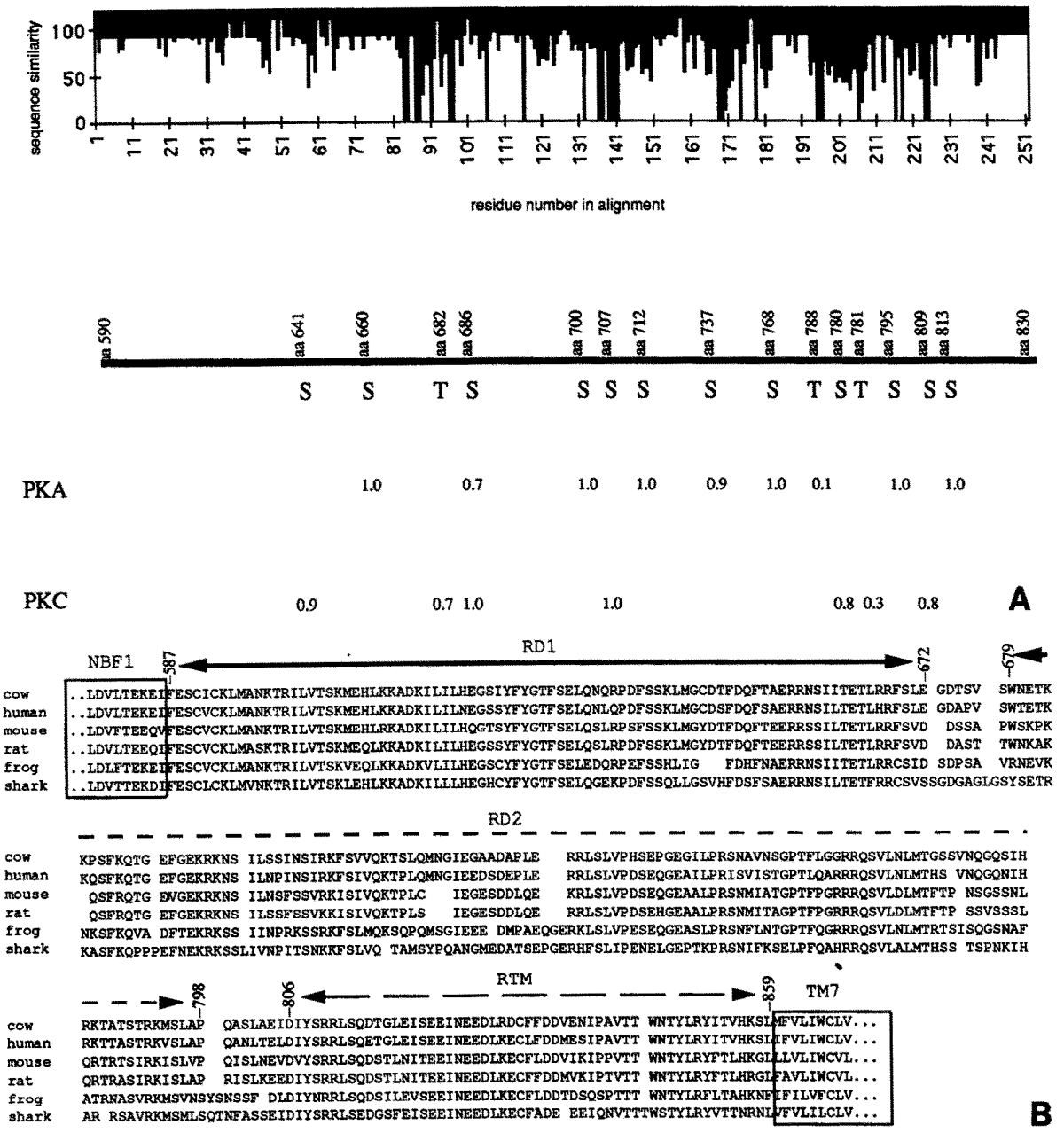
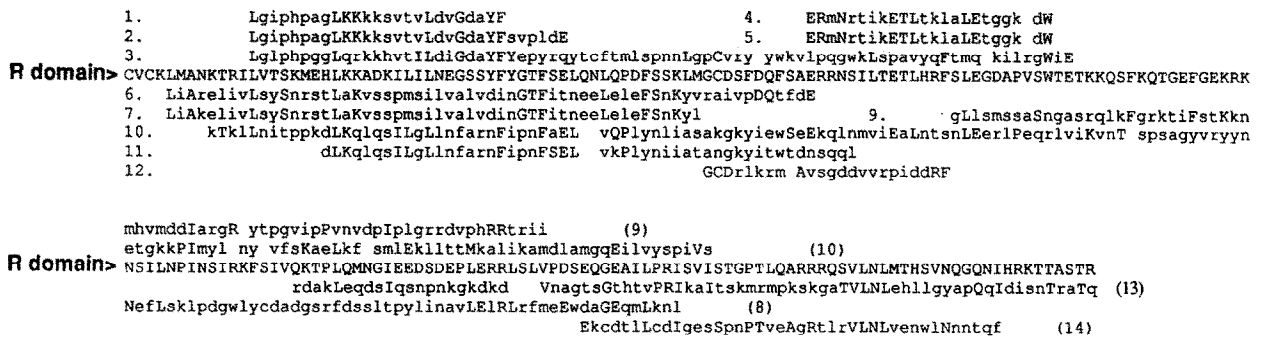


Fig. 1. Amino acid sequence conservation in the R domain. (A) Amino acid sequences from ten animal species were aligned using the SEQSEE algorithm. Each position in the alignment was given a numerical score based on the degree of conservation. Alignment position 1 corresponds to residue 590 in human CFTR and position 251 to residue 830. The fractional conservation of the putative phosphorylation sites by PKA and PKC are also indicated. (B) Sequence alignment of R domain from six species, indicating proposed domain boundaries. Abbreviations: RD1, 1st R domain domain; RD2, 2nd R domain domain; RTM, domain linking R domain and transmembrane 7; TM7, transmembrane domain 7.

that the longer the length of the alignment of two sequences, the lower the identity between the sequences needs to be for the two proteins to be related structurally [42]. This trend suggests that the R domain may be related to the pol proteins, and the relationship warranted further investigation.

It was considered that if there exists any biologically relevant sequence similarity between the viral pol proteins and the R domain, the same part of the R domain should align with the same functional part of the viral pol protein. The functional parts of the various pol proteins are not immediately obvious, as the various proteins contain different enzymatic activities. To provide functional identity of the residues in the pol proteins, each was aligned with the Moloney murine leukemia virus (MMLV) pol protein [43]. Since the common characteristic of the 14 pol proteins with putative relationship to the R domain is their similarity to RNA polymerase activities, an RNA polymerase might have been chosen as a functional reference. However, the MMLV pol was



#	virus	residues	% identity	fasta score	s.d. from mean	reference
1.	siv	272-297	32	52	5.5	28
2.	hiv	247-276	29	52	5.5	29
3.	visna	234-315	17	63	6.5	25
4.	ape leukemia	979-1002	42	70	7.6	30
5.	simian sarcoma	108-130	42	68	7.3	31
6.	bovine rotavirus	225-288	25	62	6.3	38
7.	simian rotavirus	225-278	25	71	7.8	37
8.	pea mosaic	2628-2682	17	56	5.3	36
9.	porcine rotavirus	394-463	16	58	5.7	33
10.	human foamy	179-334	20	61	6.2	26
11.	simian foamy	401-451	21	57	5.7	27
12.	yellow fever	3163-3187	21	49	4.2	35
13.	potato y	1302-1377	21	48	3.3	32
14.	Dengue	2626-2668	31	55	5.3	34

Fig. 2. Alignment of the R domain with viral pol proteins. The R domain sequence is in the center in capital letters and the viral sequences are aligned around it. Where the viral sequence shows exact conservation, the viral residue is capitalized. The table indicates features of each aligned viral sequence. s.d., standard deviations from mean score obtained searching the database.

chosen because: more of the pol proteins were retroviral, alignments for some of the pol proteins with the MMLV were available in the literature and could be used to confirm the alignments performed here, there is considerable information available about the structure/function of the MMLV pol protein and some of the aligned fragments from the retroviral proteins might code for non-polymerse activities.

The alignments of the 14 viral pol sequences with the MMLV pol obtained with the SEQSEE algorithm agreed with those reported in the literature [21,25,26,44]. Insufficient information is available for a rigorous evaluation of the goodness of the alignment. However, if the alignments are inaccurate, they will not be expected to contribute false positive conclusions in this study. Considering that, ignoring gaps, there are 2100 (approximate sum of the lengths of the MMLV pol protein and the other viral pol protein) possible alignments of a viral protein with MMLV pol. There are 13 other potential alignments of the R domain with the MMLV pol sequence, from each of the other viral proteins. It would be expected that an incorrect alignment of a viral protein would then coincide with one of the other 13 alignments by chance at a frequency of 13/2100 = 0.006. Gaps are ignored here for simplicity; if gaps are included, there are more possible alignments, and the outcome becomes more unlikely as a result of chance.

The sequence similarity score from the alignment of each viral pol protein with the R domain was calculated using the SEQSEE scoring matrix. The fragment of the viral sequence identified using the FASTA algorithm was compared to the alignment of the ten R domain sequences. The value assigned for the sequence similarity at any position was the sum of the conservation scores from the SEQSEE matrix obtained by comparing the residue in the viral sequence to the R domain residue from each species in R domain sequence alignment. Where the same R domain residue aligned with the same MMLV pol protein residue in more than one viral protein, the sequence similarity score was taken as the sum of all the scores at that position.

The sequence similarity score was plotted as a function of the residue number in the R domain and the residue number of the MMLV pol protein to which each of the viral sequences was aligned. Fig. 3A contains this plot. Significance of alignment of the R domain and the MMLV pol protein mapping was judged on two properties of the plot: the position of the lines in the R domain residue–pol protein residue plane and the magnitude of the sequence similarity score.

If the alignments have biological relevance, then the same fragment of the R domain should align with the same functional part of the pol protein. Amidst a background of noise, there is 'constructive interference' in the Fig. 3A. Constructive interference is indicated by areas on the plot where the lines are more intense (generally blacker) and higher, as a result of the coincidence of more than one data set and high sequence similarity

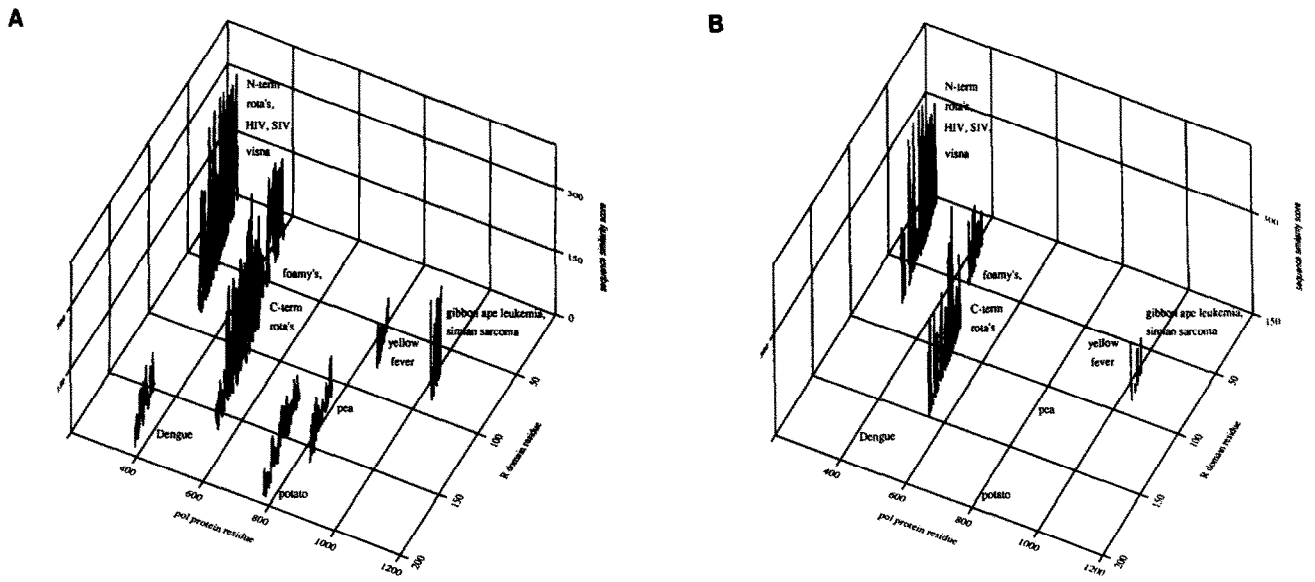


Fig. 3. Sequence similarity between the R domain and 14 viral pol proteins. The degree of sequence similarity was calculated for each alignment of the R domain and the viral protein, and then plotted as a function of the residue number in the R domain and the residue number of the MMLV pol protein to which the viral sequence was mapped. The type of virus from which each line originated is indicated on the graph. (A) Sequence similarity scores are shown starting from a value of zero. (B) The threshold for sequence similarity has been raised to a value of 150, deemed to represent conservative substitution.

score. R domain residues 15 to 70 align with equivalent pol protein residues 245–300 in HIV, SIV, visna and the N-terminal fragment of the rotaviruses. Similarly, constructive interference is observed in the alignment of R domain residues 80–150 and the MMLV pol protein residues 490 to 560 in the foamy and C-terminal fragment of the rotaviruses. It is particularly interesting that the alignment of the rotaviruses constructively interfere with those of the retroviruses. The rotaviruses and retroviruses are not generally consider to be related. However, the likelihood of the rotavirus alignments coinciding with the retroviral alignments of the R domain, particularly in two disjunct sequences by chance is low and is likely biologically relevant. Sequences which lack constructive interference, such as the Dengue, yellow fever, pea mosaic, and potato cannot be called significant by this analysis.

The magnitude of the sequence similarity was also considered. A score of 150 was deemed to signify conservative substitution and the similarity scores were plotted with this threshold, as shown in Fig. 3B. It is evident that the sequence alignment of the R domain at residues 15–70 shows higher similarity to pol residues 245–300 than to pol residues around 400 to 450. Similarly, the sequence similarity provided by the gibbon ape leukemia and simian sarcoma protein alignments around pol residues 1000 to 1030 are less significant than those around 500. Therefore, it appears that the R domain is related to the viral pol protein, with the R domain residues 15–70 and 80–150 related to, respectively, pol protein residues 245 to 300 and 490–560, in the residue numbering scheme of the MMLV pol protein.

2.6. *Relevance of the relationship between R domain and pol proteins*

The relationship between the R domain and the viral pol proteins supports the two domain model for the R domain. The proposed RD1 domain is related to part of the pol protein which is not contiguous with the part of the pol protein to which the proposed RD2 domain appears to be related. The existence of two domains within the R domain is supported by data describing the function of CFTR with deletions in the R domain [45]. Based on 'functional domains', the boundary between the first and second domain has been postulated to be at amino acid 708. It is easy to envision that to maintain a functional molecule with an internal domain deleted, a few amino acids from the deleted domain may be required to keep the remainder of the molecule in the correct spatial orientation. This may explain the discrepancy between the boundaries presented here and those proposed in [45].

A functional relationship between the R domain and the pol proteins is unlikely based on what is known of the functions of CFTR and the pol proteins. Thus, it is likely that the pol proteins and R domain share some three-dimensional characteristics in areas where the sequences are related. There exists precedent for proteins which show low sequence similarity and little functional similarity but structural relationships [2,46]. The MMLV residues 245 to 300 are located in the retroviral reverse transcriptase between the nucleotide binding site and the catalytic center [47]. MMLV residues 490–560 are located in what has been called the tether [39] region of the retroviral pol protein between the reverse transciptase

and the RNaseH. At present, there is no three-dimensional structural data available for pol proteins; however, when it becomes available it may provide insight into the structure of the R domain.

Analysis of the amino acid sequence of the R domain of CFTR has suggested that it may in fact be composed of two domains. This finding should further the knowledge of the structure/function relationship of CFTR.

Acknowledgements: We would like to thank J. Forman-Kay for her helpful comments on the manuscript and P. Romkey for his technical assistance. A.M.D. is a recipient of an MRC-ICI Pharma fellowship. This work was supported by grants from the US and Canadian Cystic Fibrosis Foundations and the NIH-NIDDK.

References

[1] Rost, B., Schneider, R. and Sander, C. (1993) Trends Biochem. Sci. 18, 120–123.
[2] Bajorath, J., Stenkamp, R. and Aruffo, A. (1993) Protein Sci. 2, 1798–1810.
[3] Rommens, J., Iannuzzi, M., Kerem, B.-S., Drumm, M., Melmer, G., Dean, M., Rozmahel, R., Cole, J., Kennedy, D., Hidaka, N., Zsiga, M., Buchwald, M., Riordan, J., Tsui, L.-C. and Collins, F. (1989) Science 245, 1059–1065.
[4] Quinton, P. (1983) Nature 301, 421–422.
[5] Welsh, M. and Liedtke, C. (1986) Nature 322, 467–470.
[6] Riordan, J., Rommens, J., Kerem, B.-S., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., Chou, J.-L., Drumm, M., Iannuzzi, M., Collins, F. and Tsui, L.-C. (1989) Science 245, 1066–1073.
[7] Eisenburg, D. (1984) Annu. Rev. Biochem. 53, 595–623.
[8] Higgins, C. (1992) Annu. Rev. Cell Biol. 8, 67–113.
[9] Cheng, S., Rich, D., Marshall, J., Gregory, R., Welsh, M. and Smith, A. (1991) Cell 66, 1027–1036.
[10] Picciotto, M., Cohn, J., Bertuzzi, G., Greengard, P. and Nairn, A. (1992) J. Biol. Chem. 267, 12742–12752.
[11] Bear, C., Li, C., Kartner, N., Bridges, R., Jensen, T., Ramjeesingh, M. and Riordan, J. (1992) Cell 68, 809–818.
[12] Rich, D., Anderson, M., Gregory, R., Cheng, S., Paul, S., Jefferson, D., McCann, J., Klinger, K., Smith, A. and Welsh, M. (1990) Nature 347, 358–363.
[13] Rich, D., Gregory, R., Anderson, M., Manavalan, P., Smith, A. and Welsh, M. (1991) Science 253, 205–207.
[14] Chang, X.-B., Tabcharani, J., Hou, Y.-X., Jensen, T., Kartner, N., Alon, N., Hanrahan, J. and Riordan, J. (1993) J. Biol. Chem. 268, 11304–11311.
[15] Wishart, D., Boyko, R., Willard, L., Richards, F. and Sykes, B. (1994) CABIOS, in press.
[16] Tata, F. Stanier, P., Wickling, C., Halford, S., Kruyer, H., Lench, N., Scambler, P., Hansen, C., Braman, J., Williamson, R. and Wainwright, B. (1991) Genomics 10, 301–307.
[17] Trezise, A., Szpirer, C. and Buchwald, M. (1992) Genomics 14, 869–874.
[18] Marshall, J., Martin, K., Picciotto, M., Hockfield, S., Nairn, A. and Kaczmarek, L. (1991) J. Biol. Chem. 266, 22749–22754.
[19] Tucker, S., Tannahill, D. and Higgins, C. (1992) Hum. Mol. Genet. 1, 77–82.
[20] Diamond, G., Scanlin, T., Zasloff, M. and Bevins, C. (1991) J. Biol. Chem. 266, 22761–22769.
[21] Doolittle, R., Feng, D.-F., Johnson, M. and McClure, M. (1989) Q. Rev. Biol. 64, 1–30.
[22] Tabcharani, J., Chang, X.-B., Riordan, J. and Hanrahan, J. (1991) Nature 352, 628–631.
[23] Berger, H., Travis, S. and Welsh, M. (1993) J. Biol. Chem. 268, 2037–2047.
[24] Pearson, W. and Lipman, D. (1988) Proc. Natl. Acad. Sci. USA 85, 2444–2448.
[25] Sonigo, P., Alizon, M., Staskus, K., Klatzmann, D., Cole, S., Danos, O., Retzel, E., Tiolain, P., Haase, A. and Wain-Hobson, S. (1985) Cell 42, 369–382.
[26] Maurer, B., Bannert, H., Daral, G. and Flugel, R. (1988) J. Virol. 62, 1590–1597.
[27] Renne, R., Friedl, E., Schweizer, M., Fleps, U., Turek, R. and Neumann-Haefelin, D. (1992) Virology 186, 597–608.
[28] Chakrabarti, L., Guyader, M., Alizon, M., Daniel, M., Desrosiers, R., Tiollais, P. and Sonigo, P. (1987) Nature 328, 543–547.
[29] Spire, B., Sire, J., Sachar, V., Rey, F., Barre-Sinoussi, F., Galibert, F., Hampe, A. and Chermann, J. (1989) Gene 81, 275–284.
[30] Delassus, S., Sonigo, P. and Wain-Hobson, S. (1989) Virology 173, 205–213.
[31] Devare, S., Reddy, E., Law, J., Robbins, K. and Aaronson, S. (1983) Proc. Natl. Acad. Sci. USA 80, 731–735.
[32] Robaglia, C. , Durand-Tardif, M., Tronchet, M., Boudazin, G., Astier-Manifacier, S. and Casse-Delbart, F. (1989) J. Gen. Virol. 70, 935–947.
[33] Johansen, E., Rasmussen, O., Heide, M. and Borkhardt, B. (1991) J. Gen. Virol. 72, 2625–2632.
[34] Hahn, Y., Galler, R., Hunkapiller, T., Dalrymple, J., Strauss, J. and Strauss, E. (1988) Virology 162, 167–180.
[35] Rice, C., Lenches, E., Eddy, S., Shin, S.J., Sheets, R. and Strauss, J. (1985) Science 229, 726–733.
[36] Fukuhara, N., Nishikawa, K., Gorziglia, M. and Kapikian, A. (1989) Virology 173, 743–749.
[37] Mitchell, D. and Both, G. (1990) Virology 177, 324–331.
[38] Cohen, J., Charpilienne, A., Chilmonczyk, S. and Estes, M. (1989) Virology 171, 131–140.
[39] Johnson, M., McClure, M., Feng, D.-F., Gray, J. and Doolittle, R. (1986) Proc. Natl. Acad. Sci. USA 83, 7648–7652.
[40] Pearson, W. (1990) Methods Enzymol. 183, 63–98.
[41] Doolittle, R. (1981) Science 214, 149–159.
[42] Sander, C. and Schneider, R. (1991) Proteins 9, 56–68.
[43] Shinnick, T., Lerner, R. and Sutcliffe, G. (1981) Nature 293, 543–548.
[44] Toh, H., Hayashida, H. and Miyata, T. (1983) Nature 305, 827–829.
[45] Rich, D., Gregory, R., Cheng, S., Smith, A. and Welch, M. (1993) 1, 221–232.
[46] Flores, T., Orengo, C., Moss, D. and Thornton, J. (1993) Protein Sci. 2, 1811–1826.
[47] Lewe, G. and Flugel, R. (1990) Virus Genes 3, 195–204.