Volume 214, number 2, 219–225 FEB 04603 April 1987

Review Letter

Studies on rationales for an expert system approach to the interpretation of protein sequence data

Preliminary analysis of the human epidermal growth factor receptor

Robert V. Fishleigh, Barry Robson, Jean Garnier* and Paul W. Finn+

Epsitron Peptide & Protein Engineering Research Unit, Department of Biochemistry & Molecular Biology, The University, Oxford Road, Manchester M13 9PT, England and **Laboratoire de Biochimie Physique, INRA, Université de Paris Sud, 91405 Orsay Cédex, France*

Received 21 January 1987

The molecular modelling of larger proteins benefits from a preliminary analysis of the sequence to identify regions of potential structural and functional importance. In this study the sequence of the epidermal growth factor receptor has been analysed using a variety of established methods and novel procedures developed for the study of weak internal and external homologies and for the use of homologous sequences in the prediction of secondary and super-secondary structures. The procedures explored here are potentially suitable for incorporation into an expert system for the initial investigation of protein sequence data.

Epidermal growth factor receptor; Insulin receptor; Structure prediction; Membrane protein; Homology; Expert system

1. INTRODUCTION

When faced with a large primary sequence implying a protein of great complexity, there is a need for a variety of sequence-analysis techniques to be applied in order to abstract the greatest amount of information before attempting detailed conformational modelling. Of particular interest are techniques which might resolve the sequence into component parts which could be treated individually in a more detailed study. Such techniques provide summary descriptions, but also naturally generate hypotheses concerning conformational roles which, when not definitive, at least represent interim and ultimately testable models. These hypothesis-generating techniques are exemplified in the present study, and represent an aspect of sequence analysis which we are attempting to incorporate into a computer-based *expert system*.

The primary sequence of epidermal growth factor receptor (EGFR) has been deduced from its cDNA [1]. As a long protein of 1210 residues, this provides an excellent opportunity for demonstrating and developing, in context, methods for determining gross conformational schemes from sequence alone.

Correspondence address: R.V. Fishleigh, Epsitron Peptide & Protein Engineering Research Unit, Department of Biochemistry & Molecular Biology, The University, Oxford Road, Manchester M13 9PT, England

+ Present address: Beecham Pharmaceuticals, Research Division, Brockham Park, Betchworth RH3 7AJ, England

2. METHODS AND RESULTS

The general procedure used for the analysis of large protein sequences can be outlined as follows. Firstly, a composition scan of the sequence is made to locate consensus sites for post-translational

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/87/$3.50

modification, identify features such as transmembrane sections and analyse regions of unusual amino acid composition. This is followed by searches for internal and external homologies, and the prediction of the secondary structures of the protein of interest and any homologous sequences identified. Information gained from the composition scan and the study of the homologues can then be used to refine the secondary structure prediction, which, in conjunction with any homology data, may permit the development of a model for the gross structure of the protein. The results of the application of this procedure to EGFR are detailed in the following sections.

An additional technique recently developed enables the division of the protein sequence into sections which may represent current or ancestral exons, collectively referred to as *para-exonic fragments*. These sections may then be studied individually in greater detail using energy minimisation and molecular dynamics. The method used is based on a statistical analysis of the distribution of nucleotide bases at exon-intron boundaries, and involves scanning the cloned DNA for sequences resembling those produced by intron excision or transcription through former introns. The non-random distribution of nucleotide bases around the mRNA splice sites leads to a bias in the amino acid residues found at the corresponding positions in the protein translation product, and hence it is also possible to search the protein sequence for regions consistent with this bias. In practice the method proved to be of only marginal benefit in the case of EGFR, and therefore a full account of the method and its application will be presented elsewhere.

2.1. *Composition scan*

An initial composition scan of the sequence identified two regions of predominantly non-polar character from positions −24 to −3 and from 622 to 644, which represent the pilot and transmembrane sequences [1]. We note that the sequence commencing at residue −4 with ASRAL and followed by the highly charged quadruplet EEKK is, on statistical grounds, a predictable site of cleavage of the signal peptide. That is, the residues A, K, S, L, R, Q, D, E and G occur commonly at cleavage sites, exemplified by the residue pairs AL (β-lactoglobulin), SR (prealbumin) and SK (*Bacillus licheniformis* penicillinase).

The concentration of 12 potential glycosylation sites (NXS, NXT) and the high content of cysteine residues in the first 622 residues of the chain would indicate that this portion of the sequence is entirely extracellular, and this is further reinforced by the abundance of potential glycosylation sites in the homologous α-subunit of the insulin receptor [2]. The intracellular region has also been attributed protein kinase activity, and indeed shows sequence homology to other protein kinases, as well as to the bulk of the *v-erb-B* oncogene product [1]. There is a potential calmodulin-dependent phosphorylation site (RXXS) in the intracellular region, but evidence from such sites should not be used injudiciously in an expert system for locating the intracellular domain, since three such sites occur before the transmembrane sequence.

The kinase relationships are under investigation by other authors (e.g. [1]) and for brevity we confine this study example to the extracellular regions of direct interest to the topic of receptor character and hormone recognition. There is, nonetheless, one interesting feature of the intracellular sequence which may emerge to be of interest in that respect. Immediately following the transmembrane sequence is the highly positively charged segment RRRHIVRKRTLRR, which is consistent with a feature interacting stably with the negatively charged phosphate head groups of the membrane lipids. This is, however, followed by segments of comparable length which alternate in overall charge, the net charge becoming progressively weaker because of the increasing inclusion of non-polar residues, and of residues of opposite charge to that dominating the segment. It is tempting to envisage that this sequence constitutes three turns of a loose spiral of 12–16 residues per turn since such a spatial arrangement would reduce electrostatic stress by the juxtapositioning of opposite charges on successive turns, and fit in well with the tendency for the occurrence of concentric layers of positive and negative salt ions near the membrane surface.

2.2. *Cysteine-rich regions*

Ullrich et al. [1] noted two cysteine-rich regions spanning residues 160–313 and 470–612, which we have designated S_1 and S_2, respectively. EGFR and the insulin receptor are not unique in containing

regions with very high densities of cysteine residues; similar concentrations may be found in such functionally distinct proteins as the low density lipoprotein (LDL) receptor, IgG3, various keratins, factors IX and X and protein C of the blood-clotting system, complement factor C9, fibronectin and high glycine-tyrosine proteins C_2 and F, for example, and this may imply general classes of conformational motifs.

The pattern of Cys residues in these regions can be analysed in terms of their separations along the sequence. In the notation used here, the separation is said to be m if Cys occurs at residue locations i and $i+m$ (fig.1). Spacings of 4, 8, 12 and 15 are seen to dominate, which is consistent with a structure with a four- or eight-residue periodicity. A corresponding repeat in hydrophobicity is observed, even if the Cys residues are not counted (not shown). It may be noted that IgG3 shows different spacings in the Cys-rich region which we suspect reflects a distinction between intra-strand and inter-strand Cys-Cys bridging, rather than a necessarily fundamental difference in architecture.

The comparison of the regions between the Cys pairs, shown in fig.2, is not intended to demonstrate homology of the fragments in isolation, since it is the sequence overall which must be considered, as in the present study. However, even at this fragment level there are indications of preferential residue types at certain positions, so that one may conceive of a consensus for a typical fragment. For example, there appears to be a consensus of CTGPGPXDC, with X a polar residue,

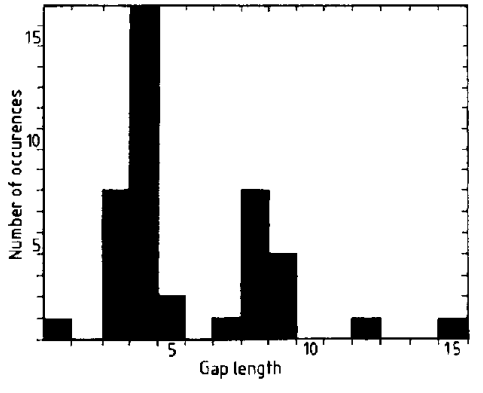

Fig.1. Histogram of inter-cysteine gaps in the extracellular portion of EGFR. The gap is m if Cys occurs at positions i and $i+m$.

(a) 163 - 3, 4, 5, 8, 8, 4, 4, 8, 1, 4, 4, 8, 3, 9, 4

267 - 4, 12, 4, 15, 3, 4, 4

475 - 7, 4, 5, 8, 3, 9, 4, 16, 3, 4, 9, 8, 3, 9, 4

593 - 3, 4, 4, 8

(b)

3	4	7
C Q K C	C D P S C	C K A T G Q V C
C L V C	C A Q Q C	
C K K C	C S G R C	8
C V S C	C H N Q C	
C I Q C	C A A G C	C Q K L T K I I C
C I Q C	C K D T C	C R G K S P S D C
C H L C	C V K K C	C T G P R E S D C
	C V R A C	C W G P E P R D C
5	C E G P C	C T G R G P D N C
	C R K V C	C T G P G L E G C
C P N G S C	C H A L C	C W G A G E E N C
C S P E G C	C V D K C	
	C H P E C	9
	C V K T C	
	C H P N C	C R K F R D E A T C
	C T Y G C	C R N V S R G R E C
		C L P Q A M N I T C
		C A H Y I D G P H C

(c) 8-3-9-4 SECTIONS

C T G P R E S D C L V C R K F R D E A T C K D T C
C W G P E P R D C V S C R N V S R G R E C V D K C
C T G R G P D N C I Q C A H Y I D G P H C V K T C

(d) 4-4-8 SECTIONS

C A Q Q C S G R C R G K S P S D C
C H N Q C A A G C T G P R E S D C
C H P N C T Y G C T G P G L E G C

Fig.2. (a) Lengths of inter-cysteine gaps in the major regions of cysteine 'repeats'. (b) Sequences bounded by cysteine residues with spacing of 3–9. (c) Homology in sections with inter-cysteine spacings of 8–3–9–4. (d) Homology in sections with inter-cysteine spacings of 4–4–8.

for the 8-separation. On this basis it is persuasive, though not fundamental to our model for EGFR, that either (i) this is convergent evolution because of the need for a particular stereoregularity, or (ii) they represent duplication of small gene fragments, presumably also leading to stereoregularity.

2.3. *Internal homology*

An internal homology scan proves to be particularly fruitful in the case of EGFR and for expert system analysis of large proteins in general. The method used for an initial analysis employed

a rule-based consideration of acceptable sequence changes primarily based on the Dayhoff matrix [3] and the similarity matrix of Levin et al. [4]. Other rules of importance include the interchangeability of P, G, A and E in loop and coil [5] and the tendency of insertion regions to favour a loop conformation [5].

This method detected a high level of homology between residues 51–144 and 367–457 (fig.3a). These regions are viewed as being potential domains and have been designated D_1 and D_2. Two shorter homologous stretches designated L_1 and L_2 were also identified and secondary structure predictions indicate that these may constitute surface loops (fig.3b).

Further regions of homology were located using a more sensitive scanning method. This latter *bit pattern* method involves a search for areas with similar patterns of hydrophobic and hydrophilic residue character. Polarity is assigned according to the propensity of a residue type to lie at the protein surface, with the consequence that non-polar residues constitute the group V, L, I, F, Y, W, M and C. This sensitive technique also demands that the order of occurrence of the short homologous fragments identified must be preserved to a certain extent, although intervening insertions are permitted.

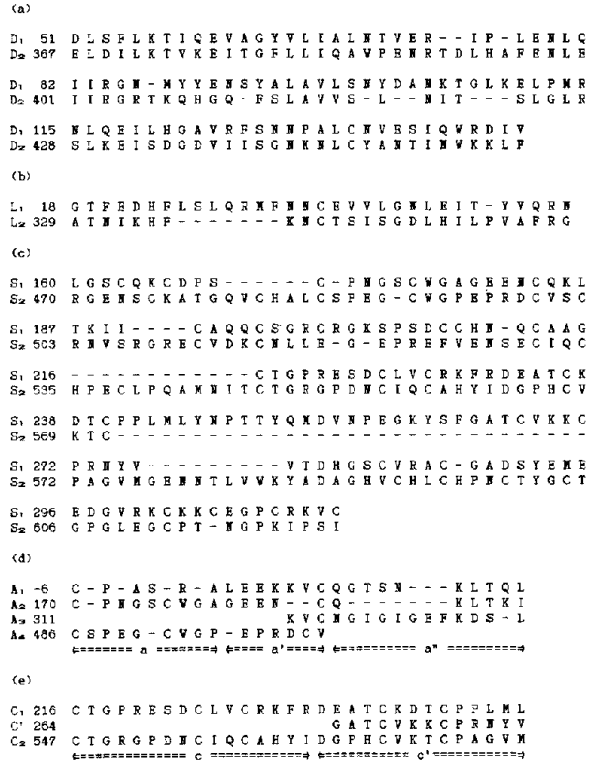

Fig.3. Alignment of homologous sections. (a) D regions, (b) L regions, (c) S regions, (d) A fragments, (e) C fragments.

Using this method an alignment was obtained for the S regions (fig.3c), and two groups of weakly homologous sections, designated A and C, were detected. The homology between the A fragments is not obvious (fig.3d). If this homology is real, then, in evolutionary terms, the A fragment would seem to be more mobile than the other fragments. In addition to an extra occurrence at the N-terminal end, the A fragment seems decomposable into a, a′ and a′′, each with some degree of independent appearance.

A major difference between the two S regions is the repetition of the short C′ fragment in the first region, separated by a 23-residue-long insertion from position 241 to 263. This insertion contains no cysteine residues and would seem to represent a discontinuity in the periodic structure of the cysteine-rich regions. It is interesting to note that the insertion of 23 residues at this point is largely compensated for by two insertions of 12 and 9 residues in the S_2 region. A comparison of the C and C′ fragments is shown in fig.3e. The relative locations of these and the other homologous regions in the sequence of EGFR are shown in a linear plot (fig.4), which also indicates the positions of features identified by the composition scan.

2.4. *Secondary structure predictions*

Secondary structure predictions, using the methods of Garnier et al. [6] and Lim [7], have been performed on the D regions. The availability of multiple sequences, coding for what presumably are essentially the same tertiary structures, enables us to apply a procedure to derive a consensus prediction of enhanced reliability (fig.5). This procedure draws upon information from the individual predictions, but also takes into consideration the location of deletion sites and the residue types present in regions of disputed conformation.

Only features of interest in the EGFR D_1 region will be discussed, but these comments can be ap-

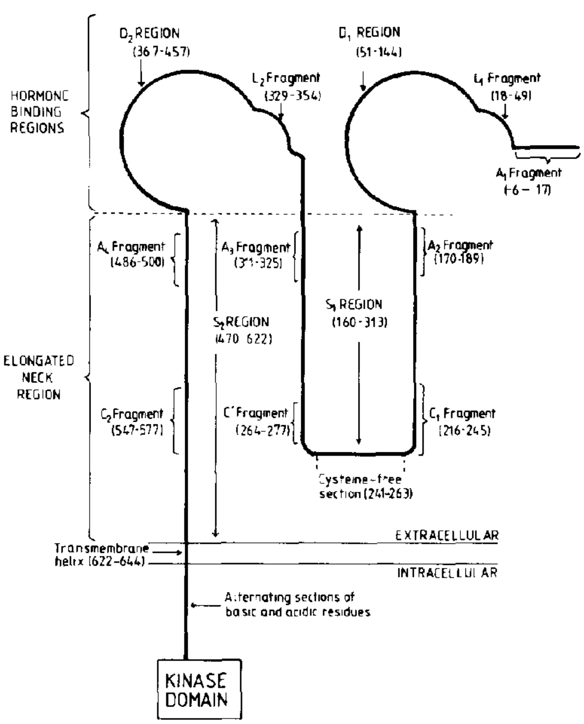

Fig.4. Relative positions of homologous regions in the EGFR sequence.

plied to the homologous sections unless otherwise stated. The section up to the start of the D regions, which is rich in insertions and deletions, contains glycosylation sites and includes the A fragments, is predicted as loop. The conserved cysteines at positions 7 and 34 could be disulphide bonded, implying a long surface loop with a possible short helix (weakly predicted around HFLSL) towards its latter half. The doublets VV and IT are predicted as extended, and may be antiparallel strands with the intervening residues in some form of loop conformation.

Extensive sections of periodic secondary structure are confined to the D regions and generally correspond to the better conserved areas. An α-helix is strongly predicted from around Asp^{51} to Gly^{63}, and is followed by a section of extended chain, which leads into a loop containing potential glycosylation sites and deletion sites. Two sections of alternating β-strand and loop follow, the latter loop bearing potential glycosylation sites in EGFR. The helical section from 113 to 120 is bounded by two predicted turns; the glycosylation site at position 111 in the insulin receptor would be on the last turn of the helix. A section of extended chain, a β-hairpin containing a conserved cysteine residue

```
              +**    *         + +        +         +* ++ * * *       +
EGFR   1  LEEKKVCQ---GT-S--N-KLTQL---GTFEDHFLSLQRMFNNCEVVLGNLEIT--YV-QRN--
EGFR 311      KVCN---GI-GIGEFKDS-LSINATNIKHF-------KNCTSISGDLHIL-PVA-FRGDS
IR     1 HLYPGEVC---PGM-DIRN-NLTRL---------HELE------NCSVIEGHLQILLM-FKTRPED
IR   309     PKVCHLLEGEKTI-D-SVTSA---------QELR------GCTVINGSLIINIRGGNNLAAE
                                                       ββ     ββ

                    +     +  + +++ +     +           + +*  ***          +
EGFR  50 Y---------DLSFLKTIQEVAGYVLIA----LNTVER--IP-LENLQIIRGN-MYYENSYALAV
EGFR 357 FTHTPPLDPQELDILKTVKEITGFLLIQ----AVPENRTDLHAFENLEIIRGRTKQHGQ-FSLAV
IR    46 FR--------DLSFPKLIM-ITDYLLLFRVYGLESLKD--L--FPNLTVIRGSRLFF-N-YALVI
IR   354 L---------EANLG-LIEEISGYLKIRRSYALVSL-S--F--FRKLRLIRGETL---E-IGNYS
                   αααααααααααα ββββββββββ               ββ          βββββ

                      + ++ *  +  + + +  +  ++ +   + +  +
EGFR  98 LSNYDANKTGLKELPMRNLQEILHGAVRFSNNPALC-NVESIQWRDIVSS....
EGFR 417 VS-L--NIT---SLGLRSLKEISDGDVIISGNKNLC-YANTINWKKLFGT....
IR    96 F---EMVH--LKELGLYNLMNITRGSVRIEKNNELC-YLATIDWSRILDS....
IR   400 F--YA-----LDNQNLRQLWDWSKHNLTITQGKLFFHYNPKLCLSEIHKM....
         β           TTTααααααααTTTβββTTTββTTTββααααααααααα
```

Fig.5. Alignment and consensus secondary structure prediction for the cysteine-poor regions of EGFR and the insulin receptor α-subunit (IR) (α, helix; β, extended chain; T, turn; blanks indicate coil). * Residue type conservation in all four sequences; $^{+}$ conservation in three. Potential glycosylation sites are underlined.

(absent in insulin receptor D_2 region) and an α-helix complete the D region. This cysteine may be involved in disulphide bridge formation with a cysteine distant in the chain.

2.5. *Possible structure of the cysteine-rich regions*

Some insight into the possible structural significance of the S regions might be obtained by comparison with other Cys-rich proteins.

In IgG3 [8,9], the Cys-rich regions represent distinct domains, with a fibrous organisation indicated by electron microscopy. Two cross disulphide bonded homologous chains are involved in forming an elongated structure, which electron microscopy has revealed as having a length of around 90 Å [8], giving an average rise per residue of 2.0 Å in this region. The high proline content of the chains would appear to eliminate α-, 3_{10} and related helices as possible structures, while the sequence repeat features would indicate the possibility of a double helix of eight residues per strand turn. The greater than average content of proline in regions S_1 and S_2 of EGFR, together with the evidence of an 8-fold repeat in hydrophobicity, *are* features in common with IgG3. One way in which the neck of IgG3 is *distinct* from EGFR is in being deficient in glycine in the Cys-rich regions.

A detected weak degree of homology exists between the N- and C-terminal sections of scale and feather keratins [10] and the S regions of EGFR (fig.6). Residue type conservation is concentrated in three sections centred around three- and four-spaced cysteines, implying similar disulphide bridging patterns, but equally important is the 'conservation' of certain intervening non-cysteine residues between EGFR and the keratins. This latter homology includes much of the region proposed by Fraser et al. [11] to be the structural region of feather keratin, in which a single polypeptide chain forms a twisted β-pleated sheet composed of four antiparallel chains of eight residues. This unexpected, if weak, homology does underline the possibility that the S regions are part of some form of related class of structural component, which, being periodic in EGFR, would seem most likely to be elongate.

This suggests a potentially testable hypothesis; that a long neck holds the hormone-binding head, or heads, away from the cell surface and resembles, at least in overall morphology, a membrane-bound immunoglobulin. In gross terms this could be tested by an electron micrographic study of the receptor in the membrane. In contrast, it has been hypothesised that the Cys-rich regions of the LDL receptor represent the binding regions, rather than purely structural features [12]. If this were so in the case of EGFR, it would place one binding domain very close to the membrane surface and would seem inconsistent with the periodic disposition of the cysteine residues.

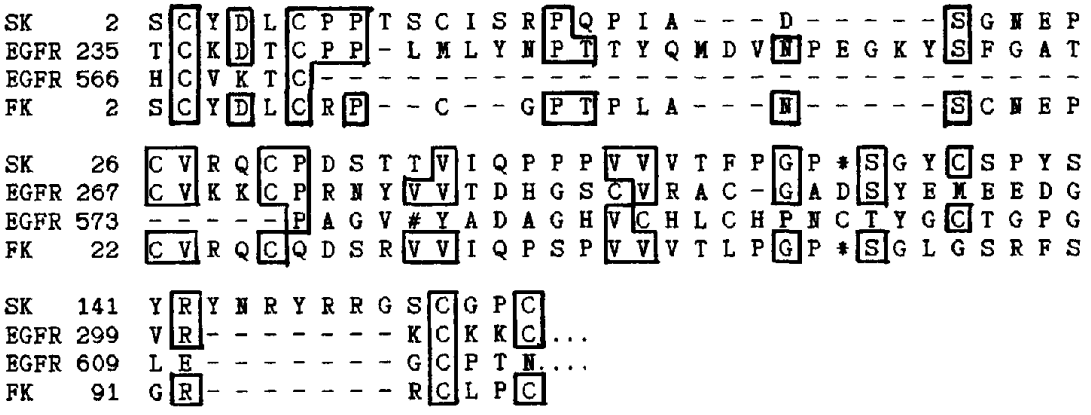

Fig.6. Alignment of scale (SK) and feather (FK) keratins with EGFR. Residues conserved between the keratins and EGFR are boxed. * Insertions of 83 and 37 residues in SK and FK. # 10-residue insertion in EGFR.

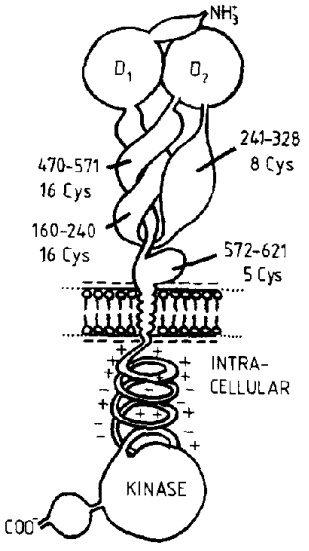

Fig.7. Model for the gross structure of EGFR.

3. CONCLUSIONS

As an example of how testable hypotheses may be generated by a preliminary analysis, fig.7 shows a model for the overall structural organisation of EGFR. The blocking of the structure into potential exonic folds only allows a limited number of solutions to the overall topography, and assuming the S regions are elongated, there is only one persuasive solution at this level of detail.

The two D regions, which we believe act as the hormone-binding component of the receptor, are held out from the cell membrane by a two-component neck structure, formed by the Cys-rich S regions. The first of these is a double-helical structure produced by interaction of residues 160–240 and 470–571, both of which contain 16 cysteine residues. The two helices are envisaged to have an 8-fold repeat with successive turns stabilised by intra-chain disulphide bridging and hydrogen bonding, while the double helix would be stabilised by inter-chain disulphide bridging.

The second component of the neck serves to link the D_2-binding head to the membrane-proximal end of the double helix, and corresponds to residues 241–328. The start of this section contains the insertion region between the duplicate copies of the short C′ fragment, which, as mentioned earlier, shows a strong tendency for a loop-like conformation. In our model we view this section as providing a flexible connector between the double helix and second half of the S_1 region. This latter unit contains eight cysteines, and may form a single internally disulphide-bonded helix, as indicated in fig.7.

The relationship of the EGFR S regions to feather keratin suggests that they may contribute to determining a specific molecular assembly, although the possibility of intra- or inter-molecular disulphide exchange, perhaps during hormone binding, may be worthy of consideration.

ACKNOWLEDGEMENTS

We thank Peter Millard for technical assistance and artwork, and Tracy Wootton for help in the preparation of the manuscript. R.V.F. is supported by the Medical Research Council.

REFERENCES

[1] Ullrich, A., Coussens, L., Hayflick, J.S., Dull, T.J., Gray, A., Tam, A.W., Lee, J., Yarden, Y., Libermann, T.A., Schlessinger, J., Downward, J., Mayes, E.L.V., Whittle, N., Waterfield, M.D. and Seeberg, P.H. (1984) Nature 309, 418–425.

[2] Ullrich, A., Bell, J.R., Chen, E.Y., Herrera, R., Petruzelli, L.M., Dull, T.J., Gray, A., Coussens, L., Liao, Y.-C., Tsubokawa, M., Mason, A., Seeburg, P.H., Grunfeld, C., Rosen, O.M. and Ramachandran, J. (1985) Nature 313, 756–761.

[3] Scwartz, R.M. and Dayhoff, M.O. (1978) in: Atlas of Protein Sequence and Structure (Dayhoff, M.O. ed.) vol.5, suppl.3, pp.353–358, National Biomedical Research Foundation, Washington, DC.

[4] Levin, J.M., Robson, B. and Garnier, J. (1986) FEBS Lett. 205, 303–308.

[5] Robson, B. and Suzuki, E. (1976) J. Mol. Biol. 107, 327–356.

[6] Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) J. Mol. Biol. 120, 97–120.

[7] Lim, V.I. (1974) J. Mol. Biol. 88, 873–894.

[8] Pumphrey, R.S.H. (1986) Immunol. Today 7, 174–178.

[9] Pumphrey, R.S.H. (1986) Immunol. Today 7, 206–211.

[10] Gregg, K., Wilton, S.D., Parry, D.A.D. and Rogers, G.E. (1984) EMBO J. 3, 175–179.

[11] Fraser, R.D.B., MacRae, T.P., Parry, D.A.D. and Suzuki, E. (1971) Polymer 12, 35–36.

[12] Yamamoto, T., Davis, C.G., Brown, M.S., Schneider, W.J., Casey, M.L., Goldstein, J.L. and Russell, D.W. (1984) Cell 39, 27–38.