Journal of Computer-Aided Molecular Design, 4 (1990) 131–145
ESCOM

J-CAMD 082

CGEMA and VGAP: a Colour Graphics Editor for Multiple Alignment using a Variable GAP penalty. Application to the muscarinic acetylcholine receptor

Henri Moereels*, Ludo De Bie and Jan P. Tollenaere

Department of Theoretical Medicinal Chemistry, Janssen Research Foundation, Turnhoutseweg 30, B-2340 Beerse, Belgium

Received 16 October 1989
Accepted 18 December 1989

Key words: Protein sequence; Alignment; Gap penalty; Colour-coding; Computer graphics; Muscarinic acetylcholine receptor; Binding site; Secondary structure

SUMMARY

Today, more than 40 protein amino acid (AA) sequences of membrane receptors coupled to guanine nucleotide binding proteins (G-proteins) are available. For those working in the field of medicinal chemistry, these sequences present a new type of information that should be taken into consideration. To make maximal use of sequence data it is essential to be able to compare different protein sequences in a similar way to that used for small molecules. A prerequisite, however, is the availability of a processing environment that enables one to handle sequences in an easy way, both by hand and by computer. In order to meet these ends, the package CGEMA (Colour Graphics Editor for Multiple Alignment) was developed in our laboratory. The programme uses a user-definable colour coding for the different AAs. Sequences can be aligned by hand or by computer, using VGAP, and both approaches can be combined. VGAP is a novel in-house written alignment programme with a variable gap penalty that also handles consecutive alignments using one sequence as a probe. In addition, secondary structure prediction tools are available.

From the 20 protein sequences, available for the muscarinic acetylcholine receptor, 13 different sequences were selected, covering the subtypes m1 to m5. By comparing the sequences, two major groups are revealed that correspond to those found by considering the transducing system coupled to the various receptor subtypes. Different parts of the protein sequences are identified as characterizing the subtype and binding the ligands, respectively.

INTRODUCTION

In general, one of the initial stages in evoking a physiological response is the formation of a drug-target complex which, in different ways, may trigger an ultimate mechanism eliciting the fi-

*To whom correspondence should be addressed.

nal response. In some cases a full three-dimensional picture is obtained from X-ray crystallography for drug–protein complexes dealing with, for example, cytochrome-P450 [1], human rhinovirus-14 [2], dihydrofolate reductase [3] and haemoglobin [4]. However, with regard to receptors, any structural information has been lacking for a long time.

Nowadays we are entering a new era in 'receptorology'. Various molecular biology techniques have provided us with the complete protein amino acid (AA) sequence information on adrenergic, muscarinic, serotonin, dopamine, peptide and light receptors. These membrane signalling receptors act by interaction with a guanine nucleotide binding protein. Today, more than 40 receptor sequences are available. For those working in the field of medicinal chemistry these AA sequences are a new challenge as they present a new type of information, that should be taken into consideration. Although we hardly understand the underlying mechanisms that govern the folding and properties of a protein, the primary structure, i.e. the AA sequence, is nevertheless a rich source of information. In fact, all the properties and characteristics of a protein, ranging from folding to function, are encoded in the AA sequence. It is a sentence written by nature. Of course, the properties may be influenced by the environment but this is only on the second level of information. In fact, it is tempting to find out how far one can get with this primary protein structure information, which is equivalent to the structural formula of a small molecule. To make maximal use of sequence data it is essential to be able to compare different protein sequences in a similar way to that used for small molecules. The use of alignment procedures to compare sequences is very popular and several computer programmes are available. A prerequisite, however, is the availability of a processing environment that enables one to handle sequences in an easy way, both by hand and by computer. In order to meet these ends the package CGEMA was developed in our laboratory and is presented in this report. In our hands the package proved to be an indispensable tool in digesting the AA sequence information which is felt to expand considerably our knowledge of how drugs work and possibly will contribute in drug design.

METHODS

In this section a description is given of the main characteristics and functions of CGEMA. It should be stressed that it has been designed as an interactive system for use on a graphics workstation. The commands covering diverse functionalities are simple and fully consistent over the whole package. Although CGEMA happens to be driven by a computer, the user is only confronted with a friendly tool and not with any computer annoyance.

CGEMA: A Colour Graphics Editor for Multiple Alignment

In the same way as structural formulae of small molecules are compared, the AA sequences of different proteins are compared by making use of alignment procedures. As the number of AAs in a protein is very large as compared with the number of chemical fragments in a small molecule, the use of a computer and computer graphics is appropriate. The comparison of two sequences is an easy task and can be performed by different alignment programmes or even by hand. The comparison of a large number of sequences is unattainable by hand within an acceptable time span. In CGEMA, sequences can be aligned by hand or by computer (using VGAP) and both approaches can be combined. The programme runs on a high-performance computer graphics workstation (Silicon Graphics, IRIS 3020 or IRIS 4D/220GTX) which is linked to a colour copy

unit (Tektronix 4692). Sequences can be entered manually or read from an external data base such as PIR (Protein Identification Resource from NBRF) and grouped under various headings covering particular interests (projects). In principle, the number of sequences and projects is unlimited and only depends on the storage capacity of the computer. In practice, a project with an alignment of 120 sequences is still easy to manage.

Aligning by hand on the graphics screen is cumbersome if one can only use a monochrome representation and the obtained alignment is hard to read. To overcome this inconvenience we introduced a coloured AA notation in CGEMA. The use of a coloured AA notation has proven to be a powerful feature in sequence comparison and allows a much faster and relaxed way of interpreting the results of sequence alignments. The AAs are colour-coded according to the following groups: Gly (G), Ala (A), Pro (P): green; Ser (S), Thr (T): orange; Asn (N), Gln (Q): pink; Asp (D), Glu (E): red; His (H), Arg (R), Lys (K): blue; Cys (C): yellow; Phe (F), Tyr (Y), Trp (W): dark blue; Met (M), Ile (I), Leu (L), Val (V): white on the screen, black on the copy. These groups correspond to those defined by the point accepted mutation matrix, which is based on AA replacements accepted by natural selection [5]. Ser and Thr, belonging to the group of Gly, Ala and Pro are given a special colour because of their polar character. Asn and Gln belong to the acidic residues Asp and Glu, but are given a different colour because they have no charge. The choice of these colours is arbitrary and can be changed at any time. Simple commands enable the user to colour each type of AA individually or according to any chosen specific characteristic like hydropathy, polarity, size or acidity. In doing so, the position of particular residues, hydrophobic or polar stretches is revealed at a glance. The use of colour should not be regarded as an artistic contribution but as a functionality enabling to disclose the results of an alignment more rapidly. In fact, in most cases, one is looking at several thousands of AA residues!

The editing and display features of CGEMA enable the user with simple commands to:

- create different projects,
- add sequences by hand or from a data bank,
- copy or delete sequences between and in different projects,
- change, insert or delete AAs and gaps,
- put numbers on and off for all or some selected sequences,
- colour each individual AA or group of selected AAs,
- change colours,
- change the order of sequences,
- colour identical, homologous and different AAs separately or in combination, relative to one particular sequence,
- create or close gaps of a certain length in all but one sequence, and
- inspect the complete alignment by moving it in a horizontal or vertical direction on the screen.

In addition, it should be stressed that the programme is designed open-ended, in order to meet the inevitable needs the user will undoubtedly express in the future.

VGAP: Alignment using a Variable GAP Penalty

In comparing sequences it is often needed to introduce gaps to optimize the correspondence between AAs. This is taken into account by introducing a gap penalty, which lowers the score indicating the goodness of fit. If the gap length is related to the penalty in a linear way (see Fig. 1), one

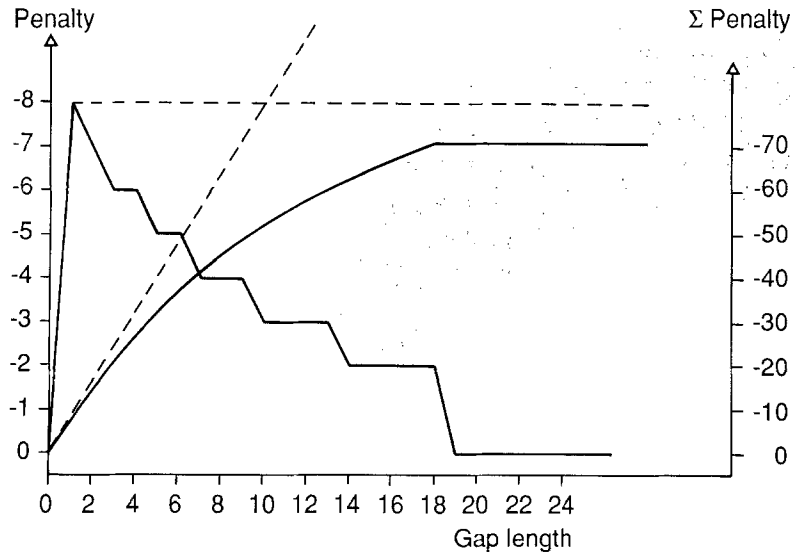


Fig. 1. Relation between penalty and gap length for a linear gap penalty function (broken line) and the non-linear gap penalty function in VGAP.

can only obtain an alignment with small gap lengths, as the penalty will overrule the final score for very large gaps. As a consequence the programme will never find a good correspondence between the sequences, which exists, for example, 150 residues down from another corresponding area. So a new alignment programme was written that uses the Needleman–Wunsch algorithm [6] but is extended with a non-linear gap penalty function. This gap penalty function is concave but never decreasing. The programme stops penalizing after a number of consecutive gaps (default: 19) and continues to introduce gaps until residues between sequences match again. The user is able to change both the penalty value and the gap length. In Fig. 2 the advantage of VGAP is illustrated schematically.

By exploiting the mathematical properties of this gap penalty function, the CPU time is reduced from $O(N^3)$ to $O(N^2 \lg N)$, while the memory requirements go from $O(N^2)$ to $O(N \lg N)$, N being the average sequence length. As a standard the programme uses the 'Point Accepted Mutation' scoring matrix of Dayhoff et al. [5]. However, different scoring matrices can be used or defined by the user. For every pair of compared sequences VGAP not only shows the optimal alignment, based on the best score obtained, but also reports the number of I (identities), H (AAs that are not identical but have a positive score in the scoring matrix) and S. The S-value, defined as the total similarity, is the sum of I and H and represents the total number of paired AAs having a positive score in the scoring matrix. The actual numbers of AAs in I, H and S are also reported in percentage relative to the shortest sequence of the compared pair. Apart from a single alignment between two sequences, VGAP also handles consecutive alignments using one sequence as a probe. The result is an alignment of n sequences in which all (n-1) sequences present in the project are aligned with the sequence indicated as the probe. If in the final alignment the order of sequences is changed by moving another sequence in the first position, the alignment remains unchanged but the scoring indices I, H and S are recalculated and interactively reported, with respect to the first sequence, which then serves as a reference.

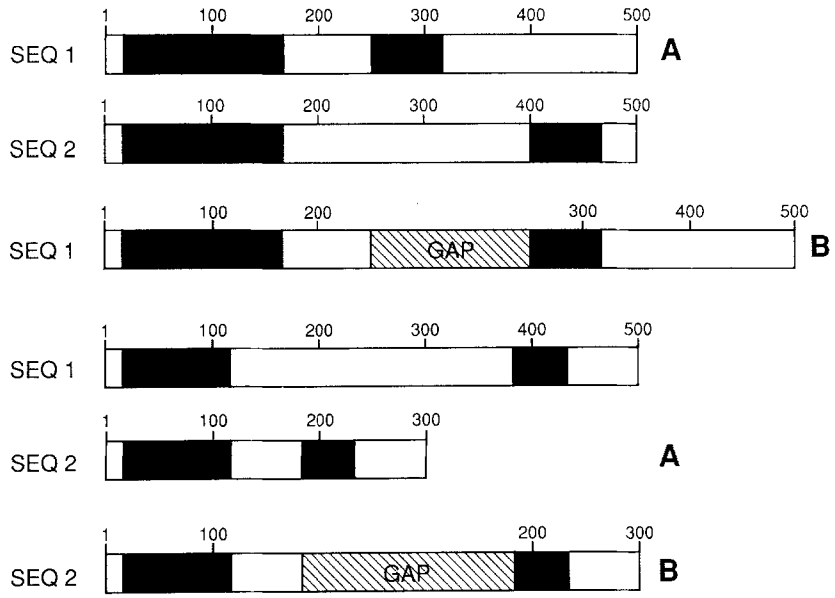


Fig. 2. Schematic representation of an alignment between sequences of equal or different length using a linear (A) or non-linear (B) gap penalty function. The dark bars indicate identical or similar areas.

The disadvantage of an alignment programme like VGAP, and many others, is that it decides on the best score obtained by the algorithm to find an optimum alignment. We frequently observed that a manual alignment has a lower score value than that obtained by a computer programme but has a higher number in I, H and S. In an alignment by hand the user can make use of intuition and experimental data available on the considered sequences. In such a guided alignment the weight of certain properties taken into account to match residues can vary as one is progressing through the sequences. For example, in certain areas one may rely on the positive score between Asp and His (according to Dayhoff) but in another area one would wish to take into account the acidic and basic properties of these residues. In CGEMA both computer-assisted and manual alignments can be combined very easily.

A very powerful tool is the command 'make table'. As a result, all sequences in a project are mutually compared and a table is constructed reporting the I- (%) and S- (%) values for each compared pair. Those sequences which are to be excluded are indicated in the same command. The sequence ordering can be changed interactively and upon introduction of a new sequence, or after correcting one, the table is updated with the same user-command but the programme only calculates the values for those pairs which did not exist yet or need to be corrected. In addition the user can ask for the alignment of any compared pair at any time. Again, the use of colour, corresponding with numerical values, greatly facilitates the inspection of elaborate tables.

The amino acid composition of proteins

The information contained in the empirical formula of a small molecule is rather primitive. If, however, we count the occurrence of particular chemical fragments, the amount of information is

much larger. In the same way, the AA composition of a protein can be a source of useful information. In fact, it is possible to classify proteins into families on the basis of their AA compositions [7]. In CGEMA the user can ask for the individual AA composition of a protein or for the contents of a user-defined group of AAs. The relative amount of hydrophobic, basic or acidic residues may in some cases give an indication of certain properties of the protein.

Tools for secondary structure prediction

The CGEMA package allows easy plotting of AA properties by either considering the individual AA positions or by moving with a window of certain residue length. For this purpose, the algorithm of Kyte and Doolittle [8] is used with different property measures like hydropathy, α-helix, β-sheet and polarity or user-defined measures. The obtained curves can be smoothed in consecutive steps and the numerical data corresponding to a particular curve are always reported in combination.

Search for patterns

In the same way as we can retrieve in a data base of small molecules to look for particular or partial structural formulae, the programme CGEMA allows one to retrieve in the user-defined projects or the PIR file. With the 'patt find' command the user can search for specific stretches of AAs or a specific pattern containing a number of undefined residues. This pattern search is very helpful to find out if for stretches or patterns of interest in a project any three-dimensional information is available. On the other hand, particular fragments or patterns with known three-dimensional structure can be looked for in proteins under study.

EXAMPLE

As an example of the use of CGEMA and VGAP a preliminary study of the muscarinic acetylcholine receptor sequences will now be presented. The muscarinic acetylcholine receptor belongs to a large family that, upon binding an appropriate hormone, transduces a signal via interaction with a G-protein (guanine nucleotide binding protein) that activates enzymatic reactions generating second messengers. The receptors interacting with G-proteins are believed to contain seven hydrophobic domains packed in a bundle of helices spanning the plasma membrane [9, 10]. In this secondary structure model, the N-terminal is located extracellularly and the C-terminal on the cytoplasmic side.

Muscarinic acetylcholine receptor sequences: Which is which?

The 20 AA sequences of muscarinic acetylcholine receptors reported in the literature are shown in Table 1. The different types are indicated with the nomenclature used by the respective authors. After mutual comparison using CGEMA and VGAP, it was found that several of these sequences are 100% identical. The human M1 sequence of Peralta et al. [12] is only different from that of Bonner et al. [9] or Allard et al. [11] in having Met instead of Val in position 173. It should be noted that Val^{173} is conserved in an Arg-Thr-Val triplet for all the other sequences. The rat M4 se-

TABLE 1
MUSCARINIC ACETYLCHOLINE RECEPTORS[a]

Human		Porcine		Rat	
m1	460 [9]	M1	460 [13]	m1	460 [9]
M1	460 [11]				
M1	460 [12]				
m2	466 [9]	M2	466 [14]	M2	466 [17, 18]
M2	466 [12]	M2	466 [15]		
m3	590 [9]	III	590 [16]	m3	589 [9]
M4	590 [12]			M4	589 [19]
m4	479 [9]			m4	478 [9]
M3	479 [12]				
m5	532 [9]			m5	531 [9]

[a]The number of AAs is indicated beside the subtype nomenclature used by the different authors.
The subtypes framed in bold have 100% identical sequences, those framed in light are only slightly different.

quence of Braun et al. [19] differs from the rat m3 of Bonner et al. [9] in that Cys^{516} and Thr^{556} are replaced by Arg^{516} and Met^{556}. Residue Cys^{516} is conserved among all muscarinic receptors while residue 556 is either Leu, His, Met or Thr. The human m3 and porcine III, which are related to rat m3, show Met. It remains an open question whether these observed differences are significant or errors, since errors seem to occur often in published sequences. Thus, it is not always easy to deal with 'different' or 'equal' results from independent experiments, especially if one also has to deal with the use of a different nomenclature indicating a specific receptor or sequence. This situation is very confusing. However, with the use of CGEMA and VGAP, differences between 'equal' sequences are detected and/or corrected very quickly, provided all the sequences of the family are in one project.

In the upper triangle of Table 2 the AA similarity between various pairs of the proteins is shown as calculated by 'make table', using the complete AA sequences. The human M1 [12] and rat M4 [19] sequences are not included as they are only slightly different from their respective congeners. In the lower triangle the mean values for the similarities between the different subtypes are given. The best similarity is observed between the sequences of the different species within a given subtype. Among the different subtypes, the largest similarity is found between m2 and m4. A high degree of similarity is also found between m1, m3 and m5. Thus, on the basis of sequence similarity, we can classify the subtypes in two groups containing (m2, m4) and (m1, m3, m5), respectively. This classification is in complete agreement with that found by considering the transducing system coupled to the various receptor subtypes. The subtypes m2 and m4 are inhibiting adenylyl cyclase, but do not activate the phosphoinositide turnover, whereas the reverse is true for m1, m3 and m5 [20].

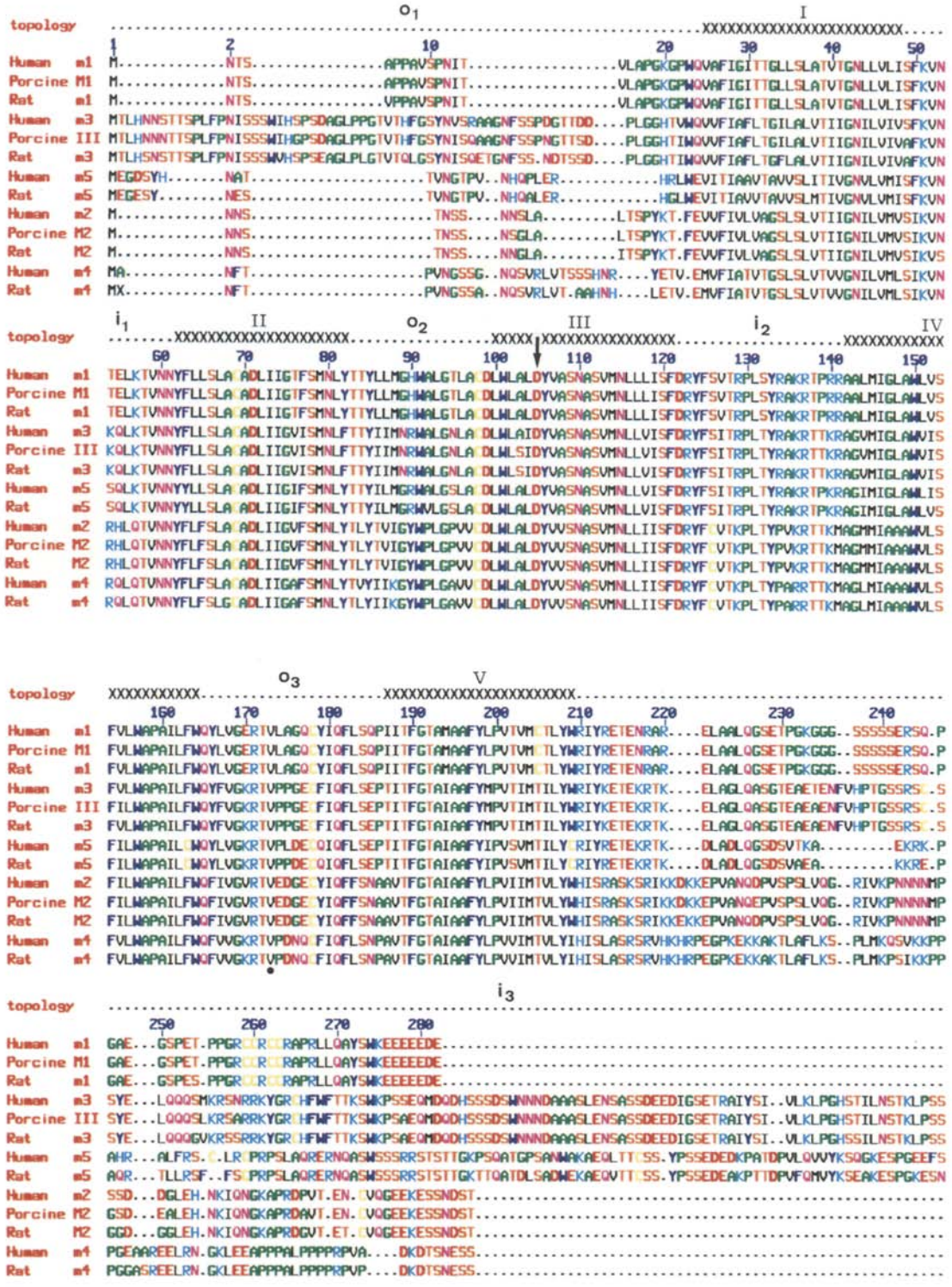
TABLE 2
AMINO ACID IDENTITY AND SIMILARITY BETWEEN VARIOUS PAIRS OF MUSCARINIC ACETYLCHOLINE RECEPTOR SEQUENCES[a]

	460 H ml H Ml	460 P Ml	460 R ml	466 H m2 H M2	466 P M2	466 R M2	590 H m3 H M4	590 P III	589 R m3	479 H m4 H M3	478 R m4	532 H m5	
531 R m5	55.4 **74.8**	54.8 **73.7**	54.6 **73.9**	45.3 **67.6**	44.0 **67.6**	43.6 **66.7**	51.6 **73.8**	52.4 **74.2**	51.2 **74.6**	48.2 **70.6**	48.5 **71.1**	89.1 **95.7**	m5
532 H m5	56.3 **75.4**	56.3 **73.9**	55.9 **74.1**	46.1 **68.2**	44.6 **68.5**	44.8 **68.0**	53.0 **75.6**	53.0 **75.2**	51.1 **73.9**	47.6 **69.5**	47.7 **69.9**		
478 R m4	47.4 **67.8**	47.0 **67.8**	47.6 **68.5**	56.7 **78.3**	58.8 **79.0**	59.2 **80.5**	47.3 **71.3**	45.6 **70.5**	46.2 **71.3**	95.8 **99.0**			
479 H m4 H M3	47.4 **67.8**	47.2 **68.0**	46.7 **67.6**	56.9 **78.5**	60.1 **80.5**	59.4 **81.1**	47.0 **71.6**	46.3 **71.4**	46.3 **71.2**			48.0 (0.4) **70.3** (0.7)	m4
589 R m3	53.7 **75.7**	53.7 **75.7**	53.9 **75.9**	46.6 **70.0**	45.9 **70.6**	45.7 **71.9**	91.9 **98.0**	93.9 **98.6**					
590 P III	54.3 **76.1**	54.8 **76.5**	53.5 **75.9**	45.7 **71.2**	45.9 **70.0**	46.1 **71.0**	95.9 **99.0**						
590 H m3 H M4	54.6 **76.7**	54.1 **76.5**	54.3 **76.7**	47.0 **70.2**	47.4 **69.3**	46.8 **68.7**		93.9 (2.0) **98.5** (0.5)		46.4 (0.6) **71.2** (0.4)		52.0 (0.9) **74.5** (0.7)	m3

TABLE 2 (continued)

466 R M2	45.7 **67.6**	44.6 **68.5**	45.7 **67.8**	**94.4** **98.3**	94.8 **98.7**				
466 P M2	43.9 **67.4**	43.9 **67.2**	43.5 **67.6**	97.4 **99.6**					
466 H m2 H M2	45.9 **69.6**	46.1 **69.3**	45.7 **69.6**		95.5 (1.6) **98.9** (0.7)	46.3 (0.6) **70.3** (1.0)	58.5 (1.4) **79.6** (1.2)	44.7 (0.9) **67.8** (0.6)	m2
460 R ml	98.7 **99.8**	97.8 **99.8**							
460 P Ml	99.1 **100**		98.5 (0.7) **99.9** (0.1)		45.0 (1.0) **68.3** (1.0)	54.1 (0.4) **76.2** (0.4)	47.2 (0.3) **67.9** (0.3)	55.5 (0.7) **74.3** (0.7)	m1
			m1		m2	m3	m4	m5	

[a]Upper number: % identity (I). Lower number (in bold): % total similarity (S). In the lower triangle: mean of the values from different species present in the compared subtypes and standard deviation between brackets.


topology
Human m1
Porcine M1
Rat m1
Human m3
Porcine III
Rat m3
Human m5
Rat m5
Human m2
Porcine M2
Rat M2
Human m4
Rat m4

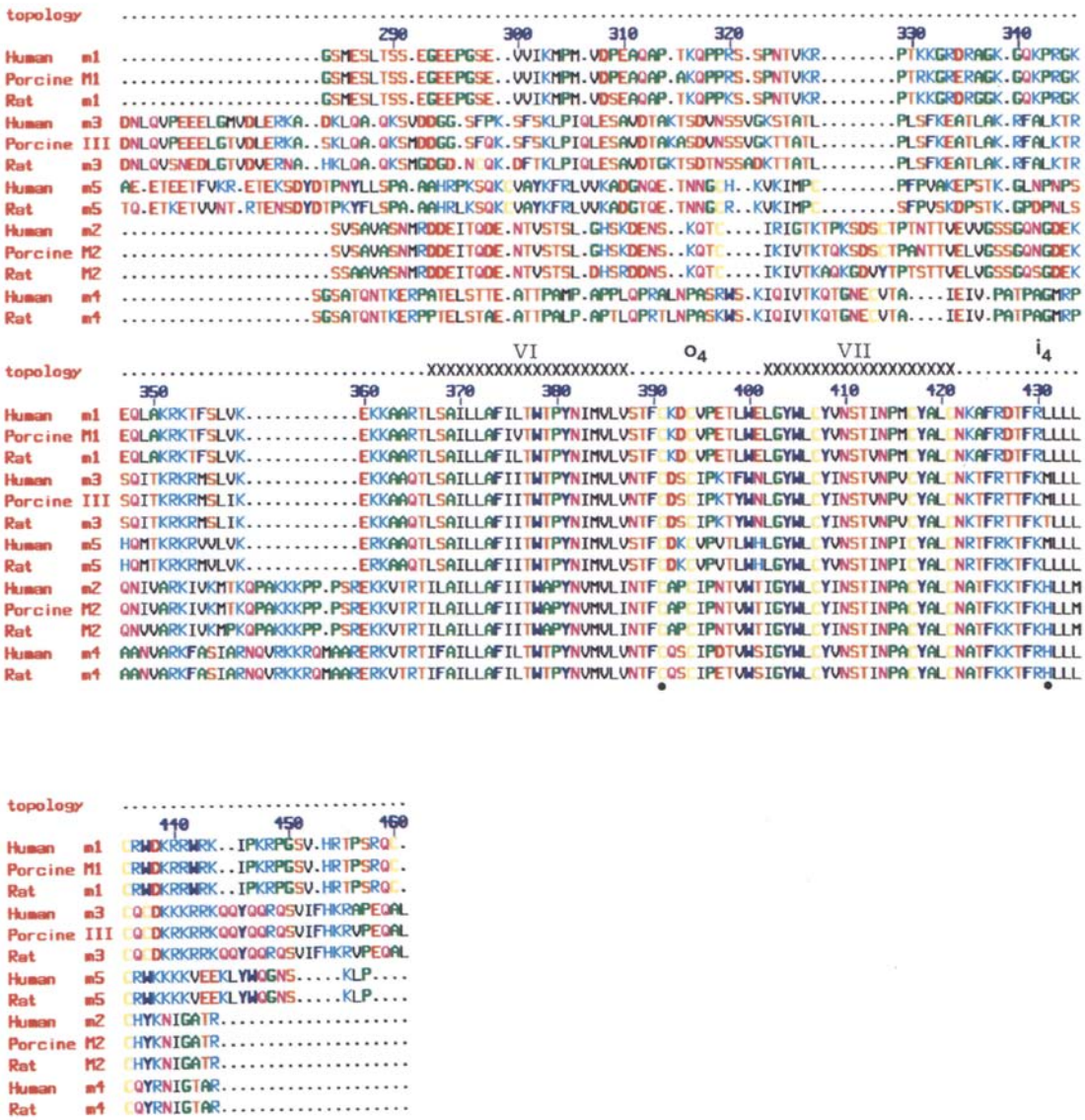


Fig. 3. Alignment of 13 selected muscarinic acetylcholine receptor sequences using a coloured one letter AA notation. The indication of the receptor topology is taken from Bonner et al. [9]. The AA numbering corresponds to the human m1 sequence. The D^{105} residue suggested to be involved in ligand binding is indicated by an arrow. ●: V^{173}, C^{516} (No. 391 in human m1) and M^{556} (No. 431 in human m1): see text.

Colour-coded alignment

Figure 3 shows an alignment, containing 6567 AA residues, of the different sequences using our standard colour-coded AA notation. According to this alignment, 158 residues are fully conserved among the different subtypes, which is 34% with respect to the shortest sequence. As can be seen in Fig. 3 and Table 3, the number of identities is largest in the transmembrane domains. Another striking feature is that considering only the cytoplasmic loops i_1, i_2 and the first 14 residues of i_4, covering 47 residues, is sufficient to discriminate between the groups m2, m4 and m1, m3, m5. This cannot be achieved by looking at the transmembrane domains only.

Identification of the ligand binding domain

Based on chemical considerations and the sequence alignment it is possible to propose a ligand binding domain, albeit without going into atomic details. With respect to ligand binding, it is reasonable to look for anionic residues in the sequence, capable of interacting with the cationic head of the ligands. According to the alignment there are four Asp residues and one Glu residue which are fully conserved in all sequences. Although one cannot exclude the presence of lipophilic antagonists in the cytoplasm (trapped molecules), we can assume that most of the molecules approach the receptor from the extracellular site. Reasoning along these lines it is unlikely that a ligand will bind in the cytoplasmic site. So the Glu residue at the end of i_3 is an unlikely candidate for ligand binding. The two Asp residues at the beginning and the end of III are unlikely as they are in the neighbourhood of the polar lipid head groups. As a consequence, the most likely candidates for ligand binding are the two Asp residues located in the second and third transmembrane domains, respectively. This hypothesis is in agreement with the experimental results identifying Asp^{113} (in III) in the β-receptor and Asp^{105} (in III) in m1 as likely candidates for interaction with the cationic ligand head [21, 22, 23].

As proposed for the β_2-adrenergic receptor [24], it can be assumed that the ligands are binding within the hydrophobic region formed by the seven transmembrane domains. These regions share a lot of similarity as shown in Table 3 and Fig. 3. However, with respect to ligand binding it is more appropriate to consider the chemical properties of the AAs rather than their homology. According to evolutionary principles Asp and Asn may be related, but from a chemical point of view they may have a different influence on binding, as is illustrated for the β_2-adrenergic receptor [21]. In Fig. 4, only the transmembrane domains are represented and the differences in AAs are indicated with respect to the human m1 sequence. If it is assumed that the cationic head of all ligands is binding to Asp in III, then it can be suggested that the other parts of the ligand molecules may interact with different residues in different domains. Thus, the observed AA differences indicated in Fig. 4 may be responsible for the differential binding of various ligand molecules. It is obvious that binding to different residues may also result in different conformational changes of the receptor, which, in turn, may influence the interaction of the cytoplasmic loops of the receptor with its second messenger.

To a large extent the above hypothesis is based on the assumption that the presumed model of the receptor, characterized by seven transmembrane domains forming a cavity in the membrane, is close to reality. However, it should be stressed that there is as yet no solid proof of the existence of the seven hydrophobic domains. The more so as the definition of these domains, based on theo-

TABLE 3

MEAN AMINO ACID IDENTITY AND SIMILARITY BETWEEN VARIOUS PAIRS OF MUSCARINIC ACETYLCHOLINE RECEPTOR SUBTYPE SEQUENCES[a]

Sequence[b]	m1 m5	m1 m3	m1 m2	m1 m4	m2 m4	m2 m3	m2 m5	m3 m5	m3 m4	m4 m5	Number of amino acids
Complete[c]	55.5 (0.7) **74.3** (0.7)	54.1 (0.4) **76.2** (0.4)	45.0 (1.0) **68.3** (1.0)	47.2 (0.3) **67.9** (0.3)	58.5 (1.4) **79.6** (1.2)	46.3 (0.6) **70.3** (1.0)	44.7 (0.9) **67.8** (0.6)	52.0 (0.9) **74.5** (0.7)	46.4 (0.6) **71.2** (0.4)	48.0 (0.4) **70.3** (0.7)	460–589
TM and the O-loops	78.6 (0.2) **91.0** (0.3)	75.5 (0.4) **92.5** (0.2)	68.9 (0.2) **87.4** (0.0)	72.2 (0.2) **87.9** (0.0)	83.6 (0.0) **95.2** (0.0)	73.2 (0.2) **90.8** (0.0)	69.1 (0.0) **88.9** (0.0)	79.3 (0.4) **95.2** (0.4)	74.1 (0.5) **90.3** (0.0)	71.7 (0.3) **90.3** (0.5)	207
TM	80.6 (0.5) **92.5** (0.4)	79.4 (0.6) **94.8** (0.0)	74.5 (0.3) **90.3** (0.0)	77.8 (0.5) **90.3** (0.0)	88.6 (0.4) **97.4** (0.0)	79.7 (0.3) **95.5 (0.0)**	75.3 (0.0) **93.2** (0.3)	80.6 (0.5) **97.2** (0.3)	76.9 (0.7) **94.2** (0.0)	76.3 (0.3) **93.2** (0.3)	154
O_2, O_3, O_4-loops	72.6 (1.0) **86.8** (0.0)	64.2 (0.0) **85.5** (0.9)	52.8 (0.0) **79.2** (0.0)	55.6 (1.0) **81.1** (0.0)	68.8 (1.0) **88.7** (0.0)	54.7 (0.0) **77.4** (0.0)	50.9 (0.0) **76.4** (1.0)	75.5 (0.0) **89.3** (1.0)	66.0 (0.0) **79.2** (0.0)	58.5 (0.0) **82.0** (1.1)	53
i_1, i_2 and 14AA of i_4-loops	79.8 (1.2) **97.9** (0.0)	77.3 (1.0) **92.9** (1.0)	58.9 (1.1) **80.9** (0.0)	63.8 (0.0) **83.0** (0.0)	88.7 (1.1) **97.9** (0.0)	66.0 (1.5) **85.1** (0.0)	67.4 (1.1) **83.0** (0.0)	88.6 (1.8) **92.9** (1.1)	68.8 (1.1) **87.2** (0.0)	70.2 (0.0) **85.1** (0.0)	47
i_3-loop	29.1 (3.0) **55.1** (2.1)	23.2 (1.7) **52.6** (0.6)	15.0 (1.0) **43.0** (2.4)	14.7 (2.5) **35.7** (4.0)	26.1 (3.8) **59.6** (3.0)	16.6 (1.6) **47.7** (3.5)	15.0 (1.6) **43.4** (1.0)	26.0 (0.7) **56.3** (0.5)	13.9 (1.1) **47.8** (1.3)	18.9 (2.1) **46.6** (1.2)	157–240

[a]Calculated as the mean of the values observed for different species present in the compared subtypes. Upper number: % identity (I); lower number (in bold): % total similarity (S). Standard deviations are between brackets.

[b]See Fig. 1, TM = transmembrane domains I to VII.

[c]Values also shown in the lower triangle of Table 2.

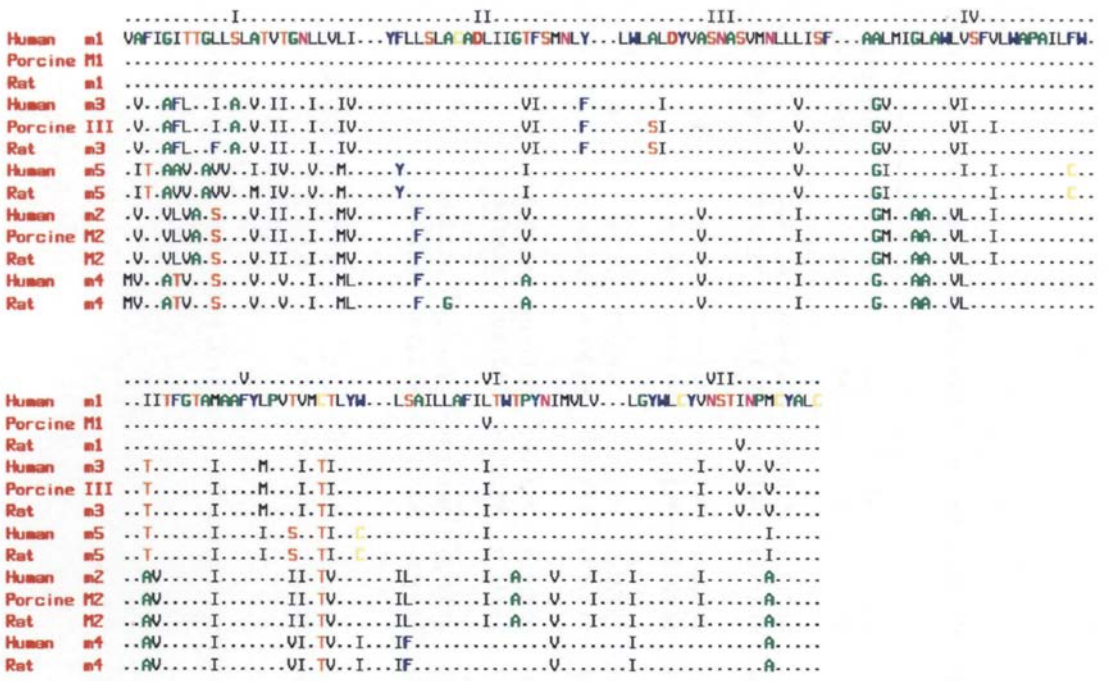


Fig. 4. Amino acids of the membrane spanning domains for 13 muscarinic acetylcholine receptor sequences. Only the AAs which are different from those observed in human m1 are indicated.

retical hydropathy plots, has not been thoroughly evaluated in a critical manner. This will be part of our further research on receptor sequences.

CONCLUSION

At this moment, 20 protein sequences are available for the muscarinic acetylcholine receptor. From this set, 13 different sequences were selected, covering the subtypes m1 to m5. By comparing the complete sequences for similarity, two major groups are revealed. One group contains the subtypes m2 and m4, which are reported to inhibit adenylyl cyclase, but do not activate the phosphoinositide turnover. The reverse is true for the subtypes m1, m3 and m5, which are found in the second homologous group. Comparison of partial sequences corresponding to the loops connecting the hydrophobic domains is sufficient, however, to characterize these groups. It is suggested that the extracellular and cytoplasmic loops characterize the subtype while the AA differences in the highly homologous hydrophobic domains are responsible for the differential ligand binding reported in the literature.

The CGEMA package and VGAP have proven to be a very useful and indispensable tool in the study of AA sequences.

Note

During the final preparation of this report two more sequences became available [25, 26]. A preliminary analysis confirms the results and conclusions of this report.

REFERENCES

1 Poulos, T.L., Finzel, B.C., Gunsalus, I.C., Wagner, G.C. and Kraut, J., J. Biol. Chem., 260 (1985) 16122.
2 Smith, T.J., Kremer, M.J., Luo, M., Vriend, G., Arnold, E., Kamer, G., Rossmann, M.G., McKinlay, M.A., Diana, G.D. and Otto, M.J., Science, 233 (1986) 1286.
3 Bolin, J.T., Filman, D.J., Matthews, D.A., Hamlin, R.C. and Kraut, J., J. Biol. Chem., 257 (1982) 13650.
4 Perutz, M.F., Fermi, G., Abraham, D.J., Poyart, C. and Bursaux, E., J. Am. Chem. Soc., 108 (1986) 1064.
5 Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C., In Dayhoff, M.O. (Ed.) Atlas of Protein Sequence and Structure, Vol. 5, Suppl. 3, 1978, p. 352.
6 Needleman, S.B. and Wunsch, C.D., J. Mol. Biol., 48 (1970) 443.
7 Cornish-Bowden, A., Biochem. J., 191 (1980) 349.
8 Kyte, J. and Doolittle, R.J., J. Mol. Biol., 157 (1982) 105.
9 Bonner, T.I., Young, A.C., Brann, M.R. and Buckley, N.J., Neuron, 1 (1988) 403.
10 Dohlman, H.G., Caron, M.G. and Lefkowitz, R.J., Biochemistry, 26 (1987) 2657.
11 Allard, W.J., Sigal, I.S. and Dixon, R.A.F., Nucl. Acid Res., 15 (1987) 10604.
12 Peralta, E.G., Ashkenazi, A., Winslow, J.W., Smith, D.H., Ramachandran, J. and Capon, D.J., EMBO J., 6 (1987) 3923.
13 Kubo, T., Fukuda, K., Mikami, A., Maeda, A., Takahashi, H., Mishina, M., Haga, T., Haga, K., Ichiyama, A., Kangawa, K., Kojima, M., Matsuo, H., Hirose, T. and Numa, S., Nature, 323 (1986) 411.
14 Kubo, T., Maeda, A., Sugimoto, K., Akiba, I., Mikami, A., Takahashi, H., Haga, T., Haga, K., Ichiyama, A., Kangawa, K., Matsuo, H., Hirose, T. and Numa, S., FEBS Lett., 209 (1986) 367.
15 Peralta, E.G., Winslow, J.W., Peterson, G.L., Smith, D.H., Ashkenazi, A., Ramachandran, J., Schimerlik, M.I. and Capon, D.J., Science, 236 (1987) 600.
16 Akiba, I., Kubo, T., Maeda, A., Bujo, H., Nakai, J., Mishina, M. and Numa, S., FEBS Lett., 235 (1988) 257.
17 Gocayne, J., Robinson, D.A., FitzGerald, M.G., Chung, F.-Z., Kerlavage, A.R., Lentes, K.-U., Lai, J., Wang, C.-D., Fraser, C.M. and Venter, J.C., Proc. Natl. Acad. Sci. U.S.A., 84 (1987) 8296.
18 Kerlavage, A.R., Fraser, C.M. and Venter, J.C., Trends Pharmacol. Sci., 8 (1987) 426.
19 Braun, T., Schofield, P.R., Shivers, B.D., Pritchett, D.B. and Seeburg, P.H., Biochem. Biophys. Res. Commun., 149 (1987) 125.
20 Barnard, E.A., Nature, 335 (1988) 301.
21 Strader, C.D., Sigal, I.S., Register, R.B., Candelore, M.R., Rands, E. and Dixon, R.A.F., Proc. Natl. Acad. Sci. U.S.A., 84 (1987) 4384.
22 Galron, R. and Sokolovsky, M., Biochem. Biophys. Res. Commun., 156 (1988) 1203.
23 Curtis, C.A.M., Wheatley, M., Bansal, S., Birdsall, N.J.M., Eveleigh, P., Pedder, E.K., Poyner, D. and Hulme, E.C., J. Biol. Chem., 264 (1989) 489.
24 Dixon, R.A.F., Sigal, I.S., Rands, E., Register, R.B., Candelore, M.R., Blake, A.D. and Strader, C.D., Nature, 326 (1987) 73.
25 Shapiro, R.A., Scherer, N.M., Habecker, B.A., Subers, E.M. and Nathanson, N.M., J. Biol. Chem., 263 (1988) 18397.
26 Liao, C.-F., Themmen, A.P.N., Joho, R., Barberis, C., Birnbaumer, M. and Birnbaumer, L., J. Biol. Chem., 264 (1989) 7328.