**SELF-ORGANIZING NEURAL NETWORKS HIGHLIGHT POSSIBLE FUNCTIONALLY RELEVANT REGIONS IN THE C-DNA CODING SEQUENCES OF G-PROTEIN COUPLED RECEPTORS.**

CASADIO R.1,2+, ARRIGO P.3, FARISELLI P.1

1,2Centro Interdipartimentale per le Ricerche Biotecnologiche (CIRB) and
2Laboratory  of Biophysics, Dept. of Biology, University of Bologna, Via Irnerio 42, I-40126 Bologna, Italy;
1e-mail: casadio@kaiser.alma.unibo.it; fax: +39-51-242576; tel. +39-51-351284;
3Istituto Circuiti Elettronici, Consiglio Nazionale delle Ricerche,  Via Opera Pia 11, I-16145 Genova;
e-mail: arrigo@ice.ge.cnr.it; fax: +39-10-6475100; tel. +39-10-6475207.

+Corresponding author

Keywords: G-protein coupled receptors, coding sequences, functional determinants, neural networks

A filter based on a set of unsupervised neural networks trained with a winner-take-all strategy is used to analyze the coding sequences of three G-protein coupled receptors, whose putative functional domains have been recently described in the literature using chimeric receptor- construction. Mapping discloses signals along the coding sequences of the G-protein coupled receptors which correlate with the putative functional domains experimentally characterized. This result and our previously reported data [1] suggest the use of the filter for predicting functional determinants in proteins starting from the coding sequence.

**1. The unsupervised classifier and the extraction of locally ordered cds fragments**

The algorithm used to analyze the coding sequences of the G-protein coupled receptors is basically a variant of a self-organizing Kohonen's feature map and it has been described before [1]. The learning procedure is a bottom-up competitive learning process. Differently from neural networks based on Hebbian learning, whereas several neurons may be activated simultaneously, in the unsupervised case only a single output neuron is active at any one time. For this reason competitive learning is particularly suited to discover statistically salient features that can be used to classify a set of input patterns included in the input vectors (X). Updating of the connection weights (W) is based on a winner-take-all strategy and only the weight vector associated to the maximally activated neuron is modified. The network consists of a two-dimensional layer of 10x10 neurons and is trained on each selected coding sequence using one codon-sliding input window of variable length (including 9, 12. 15, 18 and 21 nucleotides). At the beginning, all synaptic vector components (weights) are real numbers randomly taken from the interval [1]. Weights are reinitialized after each cds presentation. Both input patterns and synaptic vectors are normalized to unit vectors. The four nucleotide bases are coded using a CLUSTAL-like (ordinal based) input code. Significant features are then extracted from the map using a logical "AND" function between two criteria. Given a kth activated neuron and the S subset of its activating vectors, the X vector is selected when it minimizes the distance from the weight vector W (Eq.1) and it maximizes the Kullback-Leibler distance (or relative entropy, Eq.2)

(1)

(2)

where Pi, and Piare respectively the frequencies of the ith nucleotide in the input vector and in the whole coding sequence. Eq. (2) gives a measure of the information relative to the extracted pattern. After this step, contiguous nucleotide fragments which activate the same neuron are concatenated according to their relative position in the coding sequence. A final selection is then made keeping only those segments which are extracted from the nucleotide sequences by at least four out of five different runs performed with variable input window lengths. A segment accepted becomes a signal in the sequence. These segments are used to locate the correspondent segments in the protein topology, so that the set of locally ordered fragments in the coding sequences is translated into the corresponding set of protein receptor segments.

**2. Analysis of single sequences and extraction of statistically significant regions**

In order to analyze what intrinsic features are unraveled by the clustering procedure, we followed this strategy: first the coding sequence is translated into the corresponding protein and secondly the segments extracted from the cds of each protein are located along the protein sequence. By this, specific protein regions are marked along the residue sequence and the coding regions selected by the filter become signaled in the residue sequence of the protein (shown as wfilter in Fig. 1) and in its transmembrane topology (as annotated in the SWISS-PROT data base).

***2.1. The human beta 2 and beta 3 adrenergic receptors***

As an example we focus on the analysis of the cdss of two beta adrenergic receptors (human beta 2 and beta 3 adrenergic receptors) recently used to construct chimeras with the aim of characterizing molecular and structural determinants involved in desensitization and sequestration of this receptor subtype [2-3]. From the data obtained with our filter, (Fig. 1A and 1B), it appears that the two sequences contain significant features in the same protein regions: in the N terminus tail, comprising the N glicosylation sites; in the third internal (i3) loop (mainly in the N-terminal part) and in the C tail comprising the serin and treonin potential phosphorylation sites. However the region corresponding to the i2 internal loop is signaled only in the beta2 adrenergic receptor (including the flanking region in the IV transmembrane segment) differently from the beta3 sequence that contains distinguished marks in the V TMH. Beta2 is also characterized by specific signals in the I TMH, in the second extracellular loop (e2) and in the VII TMH: these regions are not marked in beta3 (compare Fig. 1A and 1B).

It is known that the two isoforms (characterized by a 45% identity [4]) are differently regulated with respect to the desensitization process, being the beta3 isoform resistant to short term agonist-promoted desensitization. Constructing chimeric beta3/beta2-adrenergic receptors, in which i2, i3 and the Ct tail of beta2 receptor where substituted in the corresponding regions of the beta3 chain, it was possible to restore fast desensitization only upon insertion of i2 in the presence of i3 and Ct. From this it was concluded that i2 plays a role which is as important as that of i3 and Ct in the fast agonist-determined desensitization process [2]. From these and other studies it appears that different regions in the proteins act as regulatory phenotypes for different processes, including activation, desensitization and sequestration [3]. It can be therefore expected that in spite of the high level of homology (which favors the construction of chimeras), the two chains should contain some intrinsic features related to this different functional specificity, especially in those molecular determinants which are involved in the different regulatory processes. According to our analysis the two chains are characterized by common features but also by characteristic regions which are different in each sequence and can be related to the functional determinants experimentally described [2]. Remarkably the i2 region contains signals only in beta2 and not in beta3. Again in beta2, motifs which are reported as potential determinants for sequestration are also included in the signals located in i2 (from D130 to F139), in the cytoplasmic end of the VII TMH (from N322 to Y326) and in the C-terminus (S355, S356, T360 and S364). As to the e2 region, it was observed that in the beta2 receptor it may play a role in ligand binding and desensitization [2].

***2.2. The rat m3 muscarinic receptor***

Another example of the performance of the method is given by the results obtained analyzing the coding sequence of the rat m3 muscarinic receptor (ACM3\_RAT in Figure 2). Studies with hybrid m2/m3 muscarinic receptors have indicated that the N-terminal 16-21 residues of the i3 loop (from R252 to T272) play an important role in determining the G-protein coupling specificity of a given muscarinic receptor subtype [4]. In the m3 receptor the specificity for Gq/11 activation is also strongly dependent on the presence of a series of single aminoacids located in the i2 domain (S168, R171, R176 and R183) and at the C-terminus of the i3 loop (A488, A489, L492 and S493) [2-6]. Remarkably most of these residues are included in the different segments obtained with the filtering procedure. Other regions in the m3 receptor which contain significant features according to our analysis are the N terminus, the IV transmembrane helix, the flanking second extracellular loop (e2), the endofacial V TMH portion and the C-terminus. Mutants of the m3 receptor in these regions have been also described with altered functional properties [4].

**3. Conclusion**

According to our clustering procedure the filter extracts from the coding sequence of each protein the most representative and locally ordered patterns within the clustered ones. By this, features common to all the vectors used to project each coding sequence into the two-dimensional map are unraveled and selected according to a measure of mutual information. This is conceptually and technically different from the common use of unsupervised methods and hidden Markov models which, after a global training phase on the whole data set, only retain characteristics common to the different sequences, mainly based on highly conserved regions [7]. Our method retains characteristics of the single sequence which according to this work correlate with putative functional regions of the protein, as indicated by experimental findings. We have previously reported that when the statistically significant features unraveled by this procedure are analyzed separately for each receptor subfamily by plotting of the density of the signals along the alignment, the emergence of the more signaled regions within a subfamily (residue identify is <= 30%) is obtained and regions of the protein topology that are both common to and characteristic of the different subfamilies can be identified [1]. A comparison with the experimental results available in the literature for the different receptor subfamilies indicated a good correlation between the regions extracted by the clustering procedure and those described as relevant functional regions [1]. Our present results validate the use of the filter to detect possible functional regions in the protein coding sequences.

**A)**

**B2AR\_HUMAN MGQPGNGSAFLLAPNRSHAPDHDVTQQRDEVWVVGMGIVMSLIVLAIVFGNVLVITAIAK**

**# TMH : HHHHHHHHHHHHHHHHHHHHHHH**

**# wfilter :\*\*\*\*\*\*\*\*\*\_\_\_\_\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**B2AR\_HUMAN FERLQTVTNYFITSLACADLVMGLAVVPFGAAHILMKMWTFGNFWCEFWTSIDVLCVTAS**

**# TMH : HHHHHHHHHHHHHHHHHHHHHHH HHHHHHHHHHHHH**

**# wfilter :\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**B2AR\_HUMAN IETLCVIAVDRYFAITSPFKYQSLLTKNKARVIILMVWIVSGLTSFLPIQMHWYRATHQE**

**# TMH :HHHHHHHHH HHHHHHHHHHHHHHHHHHHHHHH**

**# wfilter :\_\_\_\_\_\_\_\_\_\_\_\*\*\*\*\*\*\*\*\*\*\_\_\_\_\_\_\_\_\_\_\_\_\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\_\_\_\_\*\*\*\*\*\*\***

**B2AR\_HUMAN AINCYANETCCDFFTNQAYAIASSIVSFYVPLVIMVFVYSRVFQEAKRQLQKIDKSEGRF**

**# TMH : HHHHHHHHHHHHHHHHHHHHHHH**

**# wfilter :\*\*\*\*\*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**B2AR\_HUMAN HVQNLSQVEQDGRTGHGLRRSSKFCLKEHKALKTLGIIMGTFTLCWLPFFIVNIVHVIQD**

**# TMH : HHHHHHHHHHHHHHHHHHHHHHH**

**# wfilter :\_\_\_\_\_\_\*\*\*\*\*\*\*\*\*\*\*\*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

B2AR\_HUMAN NLIRKEVYILLNWIGYVNSGFNPLIYCRSPDFRIAFQELLCLRRSSLKAYGNGYSSNGNT

**# TMH : HHHHHHHHHHHHHHHHHHHHHHH**

**# wfilter :\_\_\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\*\*\*\*\*\*\***

**B2AR\_HUMAN GEQSGYHVEQEKENKLLCEDLPGTEDFVGHQGTVPSDNIDSQGRNCSTNDSLL**

**# TMH :**

**# wfilter :\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**B)**

**B3AR\_HUMAN MAPWPHENSSLAPWPDLPTLAPNTANTSGLPGVPWEAALAGALLALAVLATVGGNLLVIV**

**# TMH : HHHHHHHHHHHHHHHHHHHHHHH**

**# wfilter :\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\*\*\*\*\*\*\*\*\*\*\*\*\*\*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**B3AR\_HUMAN AIAWTPRLQTMTNVFVTSLAAADLVMGLLVVPPAATLALTGHWPLGATGCELWTSVDVLC**

**# TMH :HHH HHHHHHHHHHHHHHHHHH HHHHHHHH**

**# wfilter :\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**B3AR\_HUMAN VTASIETLCALAVDRYLAVTNPLRYGALVTKRCARTAVVLVWVVSAAVSFAPIMSQWWRV**

**# TMH :HHHHHHHHHHHHH HHHHHHHHHHHHHHHHHHHHHH**

**# wfilter :\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**B3AR\_HUMAN GADAEAQRCHSNPRCCAFASNMPYVLLSSSVSFYLPLLVMLFVYARVFVVATRQLRLLRG**

**# TMH : HHHHHHHHHHHHHHHHHHHHH**

**# wfilter :\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\*\*\*\*\*\*\*\*\*\*\*\*\*\*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**B3AR\_HUMAN ELGRFPPEESPPAPSRSLAPAPVGTCAPPEGVPACGRRPARLLPLREHRALCTLGLIMGT**

**# TMH : HHHHHHH**

**# wfilter :\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\_\_\_\_\_\_\_\_\_\_**

**B3AR\_HUMAN FTLCWLPFFLANVLRALGGPSLVPGPAFLALNWLGYANSAFNPLIYCRSPDFRSAFRRLL**

**# TMH :HHHHHHHHHHHHHH HHHHHHHHHHHHHHHHHHHH**

**# wfilter :\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**B3AR\_HUMAN CRCGRRLPPEPCAAARPALFPSGVPAARSSPAQPRLCQRLDGASWGVS**

**# TMH :**

**# wfilter :\_\_\_\_\_\_\_\_\_\_\_\*\*\*\*\*\*\*\*\*\*\*\_\_\_\*\*\*\*\*\*\*\*\*\*\_\_\_\_\_\_\_\_\_\_\_\_\_**

**C)**

**ACM3\_RAT MTLHSNSTTSPLFPNISSSWVHSPSEAGLPLGTVTQLGSYNISQETGNFSSNDTSSDPLG**

**# TMH :**

**# wfilter :\_\_\_\_\_\_\*\*\*\*\*\*\*\*\*\*\*\*\*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**ACM3\_RAT GHTIWQVVFIAFLTGFLALVTIIGNILVIVAFKVNKQLKTVNNYFLLSLACADLIIGVIS**

**# TMH : HHHHHHHHHHHHHHHHHHHHHHH HHHHHHHHHHHHHHHH**

**# wfilter :\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\_\_\_\_\_\_\_\_\_\_\_\_\_**

**ACM3\_RAT MNLFTTYIIMNRWALGNLACDLWLSIDYVASNASVMNLLVISFDRYFSITRPLTYRAKRT**

**# TMH :HHHH HHHHHHHHHHHHHHHHHHHHH**

**# wfilter :\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\*\*\*\*\*\*\*\*\*\*\*\*\***

**ACM3\_RAT TKRAGVMIGLAWVISFVLWAPAILFWQYFVGKRTVPPGECFIQFLSEPTITFGTAIAAFY**

**# TMH : HHHHHHHHHHHHHHHHHHHHHH HHHHHHHHHHH**

**# wfilter :\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**ACM3\_RAT MPVTIMTILYWRIYKETEKRTKELAGLQASGTEAEAENFVHPTGSSRSCSSYELQQQGVK**

**# TMH :HHHHHHHHHHH**

**# wfilter :\_\_\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\_\_\_\_\_\_\_\_\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**ACM3\_RAT RSSRRKYGRCHFWFTTKSWKPSAEQMDQDHSSSDSWNNNDAAASLENSASSDEEDIGSET**

**# TMH :**

**# wfilter :\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**ACM3\_RAT RAIYSIVLKLPGHSSILNSTKLPSSDNLQVSNEDLGTVDVERNAHKLQAQKSMGDGDNCQ**

**# TMH :**

**# wfilter :\_\_\_\_\_\_\_\_\_\_\_\_\*\*\*\*\*\*\*\*\*\*\_\_\_\_\_\_\_\_\_\_\_\_\*\*\*\*\*\*\*\*\*\*\*\*\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**ACM3\_RAT KDFTKLPIQLESAVDTGKTSDTNSSADKTTATLPLSFKEATLAKRFALKTRSQITKRKRM**

**# TMH :**

**# wfilter :\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**ACM3\_RAT SLIKEKKAAQTLSAILLAFIITWTPYNIMVLVNTFCDSCIPKTYWNLGYWLCYINSTVNP**

**# TMH : HHHHHHHHHHHHHHHHHHHH HHHHHHHHHHHHH**

**# wfilter :\_\_\*\*\*\*\*\*\*\*\*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*ACM3\_RAT VCYALCNKTFRTTFKTLLLCQCDKRKRRKQQYQQRQSVIFHKRVPEQAL*

**# TMH :HHHHHH**

**# wfilter :\_\_\_\_\*\*\*\*\*\*\*\*\*\_\_\_\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Figure 1 Single sequence analysis of (A) beta 2 adrenergic receptor (B2AR\_HUMAN), (B) beta 3 adrenergic receptor (B3AR\_HUMAN) and (C) rat m3 muscarinic receptor (ACM3\_RAT). TMH, Transmembrane topology (as annotated in SWISSPROT). Wfilter, signals as detected by the filtering procedure (see Text for details).

**References**

1. P. Arrigo, P. Fariselli and R. Casadio, "Self-organizing neural maps of the coding sequences of G-protein coupled receptors reveal local domains associated with potentially functional determinants in the proteins" ISMB**5**, 44 (1997)
2. R. Jokers, A. Silver, A.D. Strosberg, M. Bouvier and S. Marullo, "New molecular and structural determinants involved in beta2-adrenergic receptor desensitization and sequestration" J. Biol. Chem.**271**, 9355 (1996)
3. M. Bouvier and G. Rousseau, "Subtype-specific regulation of the beta-adrenergic receptors, Adv. Pharmacol.**42**, 433 (1998)
4. htp://swift.embl-heidelberg.de/7tm/htmls/GPCR.html
5. J. Wess, "G-protein-coupled receptors: molecular mechanisms involved in receptor activation and selectivity of G-protein recognition" FASEB J.**11**, 346 (1997)
6. N.Blin, J.Yun and J. Wess, "Mapping of single amino acid residues required for selective activation of Gq/11 by the m3 muscarinic acetylcholine receptor" J. Biol. Chem.**270**, 17741 (1995)
7. P. Baldi and Y. Chauvin,"Hidden Markov Models of the G-protein coupled receptor family" J. Comput. Biol.**1**, 311 (1994)