Self-Organizing Neural Maps of the Coding Sequences of G-Protein-Coupled Receptors Reveal Local Domains Associated with Potentially Functional Determinants in the Proteins.

Arrigo P., Fariselli P. and Casadio R.

Istituto Circuiti Elettronici, Consiglio Nazionale delle Ricerche, Via De Marini 6, I-16149 Genova;
E-mail: arrigo@ice.ge.cnr.it; fax: +39-10-6475100; tel. +39-10-6475207.
Laboratory of Biophysics, Dept of Biology, University of Bologna, Via Irnerio 42, I-40126 Bologna, Italy;
*casadio@kaiser.alma.unibo.it; fax: +39-51-242576; tel. +39-51-351284

Abstract

Mapping of the coding sequences of the best characterized subfamilies of G-protein- coupled receptors is performed with unsupervised neural networks based on a winner-take-all strategy. High order features therefrom extracted originate signals along the aligned protein sequences of the different subfamilies. These plots reveal characteristic domains common and/or characteristic of the receptor subfamily. By comparison with the existing experimental results, it is obtained that most of the regions signalled by clustering overlap with possible functional regions in the folded proteins. This is particularly noticeable for the third cytoplasmic loop, which is likely to be involved in the molecular coupling with the G-proteins. The results suggest that functional regions in proteins may be characterized by intrinsic representative features in the coding sequences which can be enlighted by high order mapping.

Introduction

In this work we describe one possible strategy for finding correlation between functional domains in proteins and the corresponding coding sequences. So far this problem has been scarcely addressed and is tackled here performing a cluster analysis of selected coding sequences with unsupervised neural networks.

Differently from supervised neural networks, the unsupervised models self-organize their activation states into topologically ordered maps (usually two-dimensional), compressing the training set of high-dimensional vectors to low-dimensional ones. The resulting maps only retain the most relevant common features of the set of input signals and are particularly suited to perform clustering of macromolecular sequences into similarity groupings (Kohonen 1995).

The method, trained with the Kohonen's unsupervised learning algorithm (Kohonen 1982), has mainly been applied to detect signal peptides (Arrigo, et al. 1991), to recognize patterns in protein sequences (Hanke, et al. 1996) and to group protein sequences into families according to their degree of sequence similarity (Ferran, et al. 1994).

A variant of the Kohonen's network model, based on a winner-take-all learning strategy, is used in this study to analyze the coding sequences of the most abundant and experimentally (functionally and structurally) well characterized subfamilies of the G-binding protein coupled receptors, a variety of cell-surface receptors which mediate their intracellular actions by a pathway that involves activation of one or more guanine nucleotide-binding regulatory proteins (G-proteins). These membrane proteins respond with a high specific interaction to different neurotransmitters and hormones, ranging from small biogenic amines to large glycoprotein hormones.

Most of the G-protein coupled receptors bear detectable sequence similarity with one another. It is commonly accepted that all share a similar topological motif consisting of seven hydrophobic well conserved alpha-helical segments that span the lipid bilayer. Moreover, a great deal of data regarding the regions involved in membrane insertion, ligand binding and coupling to G-proteins and regulatory kinases is available (Dohlman, et al. 1991; Oprian 1992; Strader, et al. 1994; Shenker 1995). Recent models also confirm a functional role of the cytoplasmic third loop for the interaction with the G-proteins (Clapham 1996).

Our results indicate that the coding regions of the different subfamilies contain local characteristic domains which correlate with specific functional regions in the receptor subfamilies. Remarkably, the region corresponding to the third cytoplasmic loop is common to all the different subfamilies.

The Unsupervised Classifier and the Extraction of Locally Ordered Domains.

The algorithm used to analyze the coding sequences of the different G protein-coupled receptors is basically a variant of a self-organizing Kohonen's feature map, previously described (Arrigo, et al. 1991). The main difference between the present and the former map is related to the updating of the connection weights. The variant is based on a winner-take-all strategy and only the weight vector associated to the maximally activated neuron is modified. The procedure simplifies the analysis of the input vectors grouped under an activated neuron.

The network consists of a two-dimensional layer of 10x10 neurons and is trained on each selected coding sequence using one codon-sliding input windows of variable length from 9 to 21 nucleotides. At the beginning, all synaptic vector components are real numbers randomly taken in the interval [0, 1]. Weights are reinitialized after each cds sequence presentation. Both input patterns and synaptic vectors are normalized to unitary vectors.

The four nucleotide bases are coded using either a Clustal-like (ordinal based) or a binary orthonormal input code. Our results are independent of the input code and routinely the ordinal code is used to speed up the network convergence.

Each input pattern is assigned to the neuron of the network that shows maximal value of activation (the winner neuron). The selected neuron has the closest synaptic vector to the input pattern, as evaluated from the Euclidean distance (Hecht-Nielsen 1990).

Then the synaptic weights of the winner neuron are modified in order to bring them closer to the vector of the input signals (X) with the following update rule:

$$W_k(t+1)=W_k(t)+\alpha(t)(W_k(t)-X) \quad (1)$$

where $\alpha(t)$ is the learning parameter $(0<\alpha(t)<1)$. $\alpha(t)$ is linearly decreased every processing cycle. For each coding sequence, the number of processing cycles is initially fixed and learning is completed when a stability criterion is satisfied (Arrigo, et al. 1991). After training has been accomplished, each input vector of the coding sequence is associated with the neuron having the closest synaptic vector. The network is simulated on a DECSTATION 5000/240 using a program written in FORTRAN 77(v.5.0)

High order features are extracted from the map using a logical "AND" function between two criteria. Given a k^{th} activated neuron and the S_K subset of its activating vectors, the X vector is selected when it minimizes the distance from the weight vector W_k (Eqn.2) and maximizes the Kullback-Leibler distance (or relative entropy (Cover and Thomas 1991)) (Eqn.3).

$$| W_k - X_k | \quad (2)$$

$$\Sigma^4_{i=1} P_i \log(P_i / \Pi_i) \quad (3)$$

where P_i, and Π_i are respectively the frequencies of the i^{th} nucleotide in the input vector and in the whole coding sequence. Eqn. (3) gives a measure of the information relative to the extracted pattern.

After this step, the nucleotide fragments are concatenated according to their relative position on the coding sequence. A further selection is then made considering only the segments extracted by all the different runs with variable input window lengths on the nucleotide sequences. These domains are used to extract the corresponding segments from the protein topology, so that the set of locally ordered domains in the coding sequences is translated into the corresponding set of protein receptor segments. Multiple sequence alignments of the protein sequences is performed using CLUSTAL VI (Higgins and Sharp 1988).

Four subfamilies of G-protein coupled receptors (adrenergic, acetylcholine (muscarinic), serotonin receptors and opsins (photoreceptors)), all belonging to the group of neurotrasmitter- and light- stimulated receptors are considered (comprising respectively 35, 16, 35 and 45 chains). Furthermore, a set of 45 protein sequences (the hormone subfamily) representative of melanocortins, glycoprotein and releasing hormones, and the so-called Family 2 receptors is also analyzed.

Results and Discussion

Patterns are extracted from all the coding sequences processed with an average of about 5 patterns per sequence. Their average nucleotide composition is quite similar to that of the initial set of coding sequences and the average length is about 7 codons per pattern..

Translating the nucleotide patterns into the corresponding protein patterns and locating them on the protein topology greatly simplifies the analysis of the filtering performance.

The frequency of signals along the sequences is plotted as a function of the protein alignment within the given subfamily. This allows a direct visualization of the regions detected by the filter and of their relative density within the subfamily. Accordingly, contiguous residues most frequently signalled within a given set, originate regions of variable length clearly emerging from the average frequency of occurrence plus one standard deviation.

When aligning the receptor sequences with one another, within a subfamily, a discernible pattern of residue conservation can be detected (data not shown). The transmembrane domains are often the most similar, whereas the N- (extracytoplasmic) and the C-(cytoplasmic) terminal regions, together with the cytoplasmic loop connecting transmembrane segments V and VI (loop V-VI) can be quite divergent.

The statistical robustness of the signal in a given position of the alignment is evaluated by a direct comparison with the density of the alignment (see for example Fig.1, where the diamonds indicate the residue density per position within the subfamily alignment). The analysis indicates that each subfamily is characterized by patterns emerging from the background signals and grouping into particular regions of the protein topology (in Fig.1 the results obtained for the adrenergic subfamily are shown). No relation is however found between the extracted patterns and the most conserved regions within the family.

Strong features common to all the subfamilies are found in the most divergent region corresponding to the V-VI loop. Moreover patterns belonging to this loop region, also extend to include portion of the V and/or the VI transmembrane regions.

Some patterns emerge as distinguished marks for some subfamilies. This is the case of patterns found in the N-terminal region of photoreceptors (including the first transmembrane segment, the first intracytoplasmic loop and the second transmembrane segment). A similar occurrence is also present in hormone receptors, whereas only the first extracytoplasmic loop is weakly signalled in serotonin receptors.

Signals in the IV transmembrane segments and flanking regions (comprising the IV extracytoplasmic and the V cytoplasmic loops) are absent in photoreceptors and present to different extents in the other subfamilies. The VII transmembrane segment (and to a much less extent its flanking regions) contains emerging features in adrenergic, serotonin and hormone receptors. Remarkably, the hormone subfamily seems to be characterized by all the pattern containing regions of the other receptors.

A list of the signalled regions in the different transmembrane helices and loops of the receptor protein models is presented in TABLE I. Most of the regions detected by our mapping procedure have also been described as relevant functional regions in the literature (TABLE I). This is so particularly for the V-VI cytoplasmic loop involved in G-protein coupling and common to all the receptors. The correlation between the local domains detected by the above procedure and functional determinants in the protein is based on the presently available experimental results (TABLE I).

In summary, a great deal of experimental results can be found in the literature pointing to the conclusion that in a protein some regions more than others seem likely to be candidate for containing functional determinants. Filtering of coding sequences of homologous proteins with unsupervised neural networks partially unravel these regions in G-protein coupled receptors and characterize them as the most representative and informative. This procedure might as well enlight those regions which starting from some common ancestral gene have been duplicated, modified and combined through evolution leading to the actual receptor kinship. It is interesting to note that the signal pattern is similar for very homologous proteins such as those grouped in the adrenergic, muscarine and serotonin receptors, whereas it differs in photoreceptors which show little homology with the previous ones.

Our classifier can therefore locate distinguished regions in coding sequences of similar proteins worth to be investigated with site-directed mutagenesis with respect to their role in protein functioning.

References

Arrigo, P.; Giuliano, F.; Scalia, F,; Rapallo, A.; and Damiani, G. 1991. Identification of a new motif on nucleic acid sequence data using Kohonen' self organizing map. *Comput. Appl. Biosci*. 7:353-357.

Dholman, H.G.; Thorner, J.; Caron M.C.; and Lefkowitz, R.J. 1991. Model systems for the study of seven-transmembrane-segment receptors. *Annu. Rev. Biochem*. 60:653-688.

Clapham, D.E. 1996. The G-protein nanomachine. *Nature* 379:297-299.

Cover, T.M.; and Thomas, J.A. 1991. *Elements of Information Theory*. Wiley & Sons, New York.

Khorana, H.G. 1992. Rhodopsin, photorecptor of the rod cell. *J. Biol. Chem*. 267:1-4.

Kobilka, B.K.; Kobilka, T.S.; Daniel, K.; Regan, J.W.; Caron, M.G.; and Lefkowitz, R.J. 1988. Chimeric α_2-, β_2-adrenergic receptors: delineation of domains involved in effector coupling and ligand binding specificity. *Science* 240:1310-1316.

Kohonen, T. 1995. *Self-organizing maps.* Springer

Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. *Biol. Cybern*. 43:59-69.

Lefkowitz, R.J. 1993. Turned on to ill-effect. *Nature* 365, 603-604.

Ferran, E.A.; Pflugfelder, B; and Ferrara, P. 1994. Self-organized neural maps of human protein sequences. *Protein Sci*. 3:507-521.

Hanke, J.; Beckmann, G.; Bork, P.; and Reich, J. 1996. Self-organizing hierarchic networks for pattern recognition in protein sequence. *Protein Sci*. 5:72-82.

Hectht-Neelsen, R. 1990. *Neurocomputing*. Addison Wesley, Reading MA (USA)

Higgins, D. A.; and Sharp, P.M. 1988. CLUSTAL: a package for performing multiple sequence alignments on a microcomputer. *Gene* 73:237-244.

Mizobe, T.; Maze, M.; Lam, V.; Suryanarayana, S.; and Kobilka, B.K. 1996. Arrangement of transmembrane domains in adrenergic receptors. *J. Biol. Chem*. 271:2387-2389.

Oprian, D.D. 1992. The ligand-binding domain of rhodopsin and other G protein-linked receptors. *J. Bioenerg. Biomembr*. 24:211-217.

Savarese, T. M; and Fraser, C.M. 1992. In vitro mutagenesis and search for the structure-function relationships among G protein-coupled receptors. *Biochem. J*. 283:1-19.

Shenker, A. 1995. G protein-coupled receptor structure and function: the impact of disease causing mutations. *Baill. Clin. Endocrin. Met*. 9:427-451.

Strader, C.D.; Fong, M.T.; Tota, M.R.; Underwood, D.; and Dixon, R.A.F. 1994. Structure and function of G protein-coupled receptors. *Annu. Rev. Biochem*. 63:101-132.

Strosberg, A.D. 1993. Structure, function and regulation of adrenergic receptors. *Protein Sci*. 2:1198-1209.

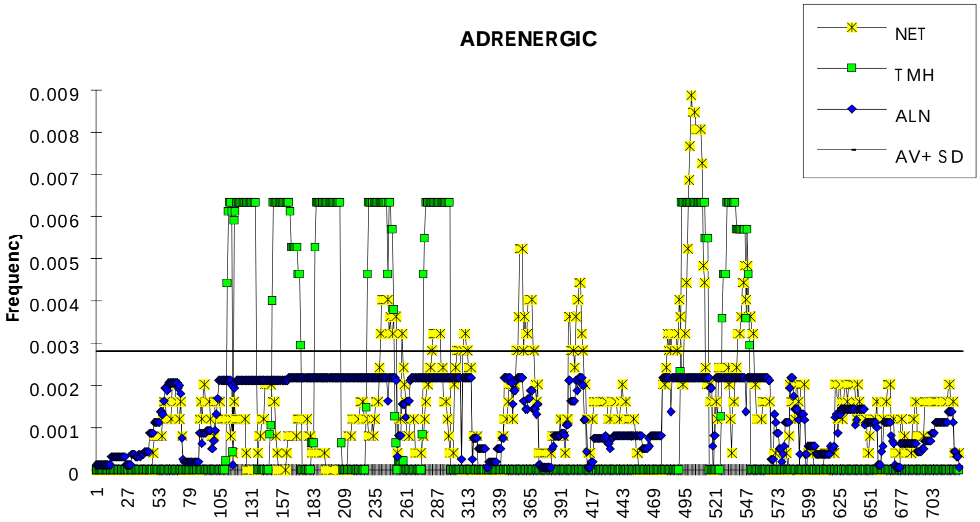

Fig.1. Frequency of high order features in adrenergic receptors along the protein alignment and topology. NET= network signal. TMH= transmembrane helices. ALN= alignment density. AV+SD= average value plus one standard deviation of the network signal.

TABLE I. Comparison between the domains extracted with the unsupervised networks and the functional determinants in G-protein-coupled receptors

	TOPOLOGICAL REGIONS														
SUBFAMILY	**1L**	**1H**	**2L**	**2H**	**3L**	**3H**	**4L**	**4H**	**5L**	**5H**	**6L**	**6H**	**7L**	**7H**	**8L**
ADRENERGIC								*°	*	*°	*°	*°	*°		
MUSCARINIC	*						*	*	*°						*
SEROTONIN	*						*	*	*	*°	*	*		*°	*
PHOTORECEPTORS	*°	*°	*°	*°			°	*	*°			*		°	°
HORMONES	*°	*	*	*°			*	*°		*°	*°	*°	*	*	*

*** Domains in which signals are above the average frequency of occurence plus one standard deviation. Transmembrane alpha helixes (H) and the interconnecting loops (L) are progressively numbered. According to the most accepted 7-helix transmembrane topology the N- and C-terminus are extracytoplasmic and intracytoplasmic, respectively.**
° Functional regions characterized with different experimental approaches as described in the literature, including hybrid construction, deletion and site-directed mutagenesis (see for review: Dholman, et al. 1991 (general); Khorana 1992 (rhodopsin); Savarese and Fraser 1992 (general); Oprian 1992 (general); Strosberg 1993 (adrenergic); Lefkowitz 1993 (hormones); Strader, et al. 1994 (general); Shenker 1995 (general).