A Weighted Cramér's V Index for the Assessment of Stability in the Fuzzy Clustering of Class C G Protein-Coupled Receptors

Alfredo Vellido*, Christiana Halka, and Àngela Nebot

Department of Computer Science, Universitat Politècnica de Catalunya,
Barcelona 08034, Spain
{avellido,christiana.halka,angela}@cs.upc.edu
http://www.cs.upc.edu/~avellido

Abstract. After decades of intensive use, K-Means is still a common choice for crisp data clustering in real-world applications, particularly in biomedicine and bioinformatics. It is well-known that different initializations of the algorithm can lead to different solutions, precluding replicability. It has also been reported that even solutions with very similar errors may widely differ. A criterion for the choice of clustering solutions according to a combination of error and stability measures has recently been suggested. It is based on the use of Cramér's V index, calculated from contingency tables, which is valid only for crisp clustering. Here, this criterion is extended to fuzzy and probabilistic clustering by first defining weighted contingency tables and a corresponding weighted Cramér's V index. The proposed method is illustrated using Fuzzy C-Means in a proteomics problem.

Keywords: Fuzzy clustering; K-Means; Clustering stability analysis; Cramér's V index; G Protein-Coupled Receptors

1 Introduction

G protein-coupled receptors (GPCRs) are cell membrane proteins of interest due to their role in transducing extracellular signals after specific ligand binding. They have in fact become a core interest for the pharmaceutical industry, as they are targets for more than a third of approved drugs [1].

GPCR functionality is mostly investigated from its crystal 3-D structure. Finding such structure, though, is a difficult undertaking[1] and only in the last decade, a handful of GPCR structures has been found, most of them belonging to the class A of the GPCR superfamily [2]. Only in 2013, one receptor not belonging to class A but to the Frizzled class and two class B [3, 4] were reported.

* This research was partially funded by Spanish MINECO TIN2012-31377 research project.

[1] Rhodopsin was the first GPCR crystal structure to be determined, back in 2000.

The first structures of the 7-trans-membrane (7TM) domains of two class C receptors were published in 2014 [5, 6].

Our research focuses on class C GPCRs, which have become a key target for new therapies [7]. The alternative approach when the tertiary 3D crystal structure is not available, is the investigation of the receptor functionality from its primary structure, that is, directly from the amino acid (AA) sequences. The comparative exploration of the sequences of the seven different described subtypes of this class may constitute a first step in the study of the molecular processes involved in receptor signalling.

Most of the existing data-based research on primary receptor sequences resorts to their alignment [8], which enables the use of more conventional quantitative analysis techniques. Given that the length of the class C GPCR sequences varies from a few hundred AAs to well over a thousand, alignment risks the loss of relevant sequence information. Alternatively, as in this paper, we can resort to methods for the analysis of alignment-free sequences from their transformation according to the AA properties (for a review see [9]).

Previous exploration of the class C GPCR sequences through visualization-oriented clustering [10] and semi-supervised analysis [11] has shown that the existing formal characterization of this class into seven subtypes only partially corresponds to the natural data cluster structure according to unaligned sequence transformations. In the current study, we investigate this issue within a more general clustering framework.

Clustering analysis often works by assigning individual data instances to one out of several clusters according to their similarity to representative examples Such assignment is often of a dichotomous or *crisp* nature: the instance either belongs to or does not belong to a given cluster. Fuzzy and probabilistic clustering methods, instead, assign each data instance to each cluster with an estimated degree or probability of membership. As a result, the uncertainty of the assignment decision is explicitly taken into account in the model.

Over the last decades [12], K-Means has become a stalwart method for data clustering, spawning many variants while remaining a common choice, even if as a benchmark, in many real-world applications. K-Means is based on crisp cluster assignments, although variants such as Fuzzy C-Means (FCM)[13] have extended the model to account for partial degrees of membership. K-Means limitations are well studied and include the lack of a closed criterion for the choice of the number of clusters K and the fact that, under different initializations, the algorithm may yield very different solutions.

Recent experimental evidence [14] has shown that K-Means solutions that might be expected to be similar according to the final value of the objective function may in fact be quite dissimilar, and that this effect increases with the value of K. This suggests the convenience of using the objective function as a criterion of model optimality only in combination with some cluster stability criterion if we aim to achieve cluster partition reproducibility. One such combined criterion is the *Separation and Concordance* (SeCo) map, which joins the standard sum-of-squares (SSQ) error and Cramér's V stability index, a variation of Pearson's χ^2, which can also be used to inform the choice of K.

The calculation of Cramér's V index is based on the use of contingency tables, which are only suitable for crisp cluster assignments. In this study, we extend the SeCo criterion to fuzzy and probabilistic clustering by first defining weighted contingency tables and a corresponding weighted Cramér's V index. This should be a more faithful assessment of the clustering solution stability for fuzzy and probabilistic methods.

The proposed methods are employed to investigate a class C GPCR primary sequence data set extracted from a publicly available database. Two experimental settings for the clustering experiments are used. The first fixes the number of clusters to the number of formal subtypes in the class in order to investigate the level of correspondence between both, while the second relaxes this constraint in order to analyze the stability of the clustering solutions using SeCo maps.

2 Methods

2.1 Alignment-Free Sequence Transformation Methods

As previously mentioned, in this paper we resort to methods for the analysis of alignment-free sequences from their transformation according to the AA properties. Three transformations were used in our experiments:

– *Amino Acid Composition (AAC)*: This simple transformation reflects the AA composition of the primary sequence. The frequencies of the 20 sequence-constituting AAs are calculated for each sequence and, as a result, an $N \times 20$ data matrix is obtained, where N is the number of data instances.
– *Auto Cross Covariance (ACC)*: The ACC transformation aims to capture the correlation of the physico-chemical AA descriptors along the sequence. The method relies on the translation of the sequences into vectors based on the principal physicochemical properties of the AAs. Data are transformed into a uniform matrix by applying a modified autocross-covariance transform [15]. First, the physico-chemical properties are represented by means of the five z-scores of AA as described in [16]. Then the Auto Covariance (AC) and Cross Covariance (CC) are computed on this first transformation. They, in turn, measure the correlation of the same descriptor (AC) or the correlation of two different descriptors (CC) between two residues separated by a lag along the sequence. From these, the ACC fixed length vectors can be obtained by concatenating the AC and CC terms for each lag up to a maximum lag, l. This transformation generates an $N \times (z^2 \cdot l)$ matrix, where $z = 5$ is the number of descriptors.
– *Digram Transformation*: The digram transformation is a particular instance of the more general n-gram transformation. It considers the frequencies of occurrence of any given pair of AAs. The n-gram concept has previously been used in protein analysis [17]. This particular transformation generates an $N \times 400$ matrix.

2.2 Data Clustering Using K-Means and Fuzzy C-Means

The K-Means algorithm creates a partition of a set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ of observed data into a set of K clusters $\Gamma = \{\Gamma_1, \ldots, \Gamma_K\}$ by defining a fixed number K of data centroids or prototypes $\{\mu_1, \ldots, \mu_K\}$. It does so by assigning individual data observations to their closest prototypes according to a given similarity measure or distance. Such assignment is crisp in the sense that individual observations are associated to individual clusters with complete certainty. Fuzzy and probabilistic clustering techniques relax this approach by assigning, to each observation, a fuzzy degree (for instance in FCM) or a probability (for instance in mixture models) of membership to each cluster. Most approaches to K-Means opt for different forms of random initialisation of the prototypes (*seeds*). The algorithm's objective is finding the set Γ that minimises a SSQ error in the form $\sum_{k=1}^{K} \sum_{\mathbf{x} \in \Gamma_k} \|\mathbf{x} - \mu_k\|^2$. FCM generalizes this objective function to become:

$$\sum_{c=1}^{C} \sum_{n=1}^{N} \omega_{nc}^{m} \|\mathbf{x}_n - \mu_c\|^2 \tag{1}$$

for C fuzzy clusters, fuzzy weights ω and fuzziness parameter m. It is well known that different initialisations of the algorithm make it converge to different local minima and that there is no guarantee of convergence to a global minimum of the objective function. In practical applications, K-Means and FCM are run with a sufficient number of random initialisations and the Γ generating minimum error is chosen among the rest.

2.3 Clustering Stability Measures

Even if finding a minimum error Γ is a central objective of K-Means, the *stability* of the clustering solution is also relevant. Solutions that are reproducible are required. That is, cluster partitions that do not change (much) under different initialisations, i.e., that are stable. This cluster stability is paramount in practical applications and can be quantified using different indices [18].

It was recently brought into attention [14] that K-Means partitions with similar errors might be greatly different from each other (thus unstable) and that this effect increases with the value of K. Assuming that solutions that strike a balance between low error and high stability ought to be sought, Lisboa *et al.* [14] proposed a framework based on the calculation of Separation/Concordance (SeCo) maps for settings using multiple random initializations of K-Means for different values of K. This entails the simultaneous display of a pair of values for each run of the algorithm, namely: The ΔSSQ, calculated as the total SSQ minus the within-cluster SSQ:

$$\sum_{k=1}^{K} \sum_{n=1}^{N} \|\mathbf{x_n} - \mu_k\|^2 - \sum_{k=1}^{K} \sum_{\mathbf{x_n} \in \Gamma_k} \|\mathbf{x_n} - \mu_k\|^2 \tag{2}$$

and a *concordance index* (CI) quantifying stability. In [14], the use of Cramér's V index is recommended as a basis for it. The CI is calculated as the median of the $(nin - 1)$ pairwise Cramér's V calculations for nin initializations. For two cluster partitions Γ and Γ' of, in turn, K and K' clusters, Cramér's V index is a variation of a χ^2 test, calculated as $V = \sqrt{\chi^2/N \centerdot min(K-1, K'-1)}$, where

$$\chi^2 = \sum_{k=1}^{K} \sum_{k'=1}^{K'} (O_{kk'} - E_{kk'})^2 / E_{kk'} \tag{3}$$

Here, $\mathbf{O}$ is an observed contingency table $(K \times K')$ matrix, whose values $O_{kk'}$ indicate the number of instances in $\mathbf{X}$ that have been assigned to cluster k in one run of the algorithm and to cluster k' in another run. The $K \times K'$ matrix $\mathbf{E}$ contains the corresponding expected values for independent cluster allocations, calculated as $E_{kk'} = \frac{1}{N}(\sum_{j=1}^{K'} O_{kj} \sum_{i=1}^{K} O_{ik'})$.

This use of contingency tables is suitable for crisp cluster assignments such as those provided by K-Means. For *soft* assignments such as those provided by FCM or Gaussian Mixture Models, instead, this use occludes the richness of the cluster solution by requiring the assignment of instances to clusters to be based on the highest degree of membership or probability.

In this paper, we propose a variation of contingency tables that better suits the characteristics of fuzzy and probabilistic models. Elements in what we call weighted observed (w$\mathbf{O}$) contingency tables will now be calculated, following the notation of Eq.1 for FCM, as $wO_{cc'} = sum_{n=1}^{N} \omega_{nc} \omega_{nc'}$; this is, for data instance n, the product of the degree of membership to cluster c in a first run of the algorithm and the degree of membership to cluster c' in a second run. Consequently, we can obtain a weighted expected (w$\mathbf{E}$) contingency table matrix whose elements are defined as $wE_{cc'} = (\sum_{j=1}^{C'} wO_{cj} \sum_{i=1}^{C} wO_{ic'})/N$. This leads to the definition of a new *weighted Cramér's V index*, where $\mathbf{O}$ is replaced by $w\mathbf{O}$ and $\mathbf{E}$ by $w\mathbf{E}$ in the calculation of χ^2 in Eq.3.

If FCM estimated that all instances had a degree of membership of 1 for a single cluster, the weighted Cramér's V index would reduce to its standard formulation. This is unlikely to happen, which means that the proposed index will lead to lower levels of CI in SeCo. This should, therefore, be not only a conservative concordance estimator, but also a more reliable clustering assessment tool, capable of distinguishing solutions with varying levels certainty.

Note that SeCo can be used as a flexible tool to chose adequate values of the K parameter (number of clusters). This is equally true when using the modified index, but, in this case, there should be no bias in favour of "over-optimistic" solutions.

3 Experiments

3.1 Materials and Experimental Setting

The data in the following experiments were extracted from GPCRDB[2] [19] (version 11.3.4 as of March 2011), a public database of G Protein-Coupled Receptor

[2] `http://www.gpcr.org/7tm`

(GPCR) protein primary sequences. The data set comprises a total of 1,510 class C GPCR sequences, belonging to seven subfamilies and including: 351 metabotropic glutamate (mG), 48 calcium sensing (CS), 208 GABA-B (GB), 344 vomeronasal (VN), 392 pheromone (Ph), 102 odorant (Od) and 65 taste (Ta). Their AA lengths vary from 250 to 1,995.

Previous research [20] investigated the supervised classification of these data sequences, from several of their alignment-free transformations, including AAC, digram, and ACC, among others. Here, we use K-Means and FCM to investigate to what extent the natural clustering structure of the data fits the subfamilies (classes) description. For that, we first report the results of experiments in which the number of clusters is fixed *a priori* to be the same as the number of class C GPCR subtypes. These will provide us with a preliminary evaluation of the level of natural subtype overlapping. We then proceed to relax that constraint and the SeCo framework, with the proposed modification of the concordance index in the case of FCM, is applied in a setting with 500 random initializations of the algorithms for different number of clusters. Following [14], only the best 10% $\triangle SSQ$ results are displayed in the SeCo plots.

3.2 Results and Discussion

The three transformed data sets were fed to the FCM and K-means algorithms. The class (subtype) specificity for each cluster for each data set was measured and the results are provided in the following paragraphs along with class-entropy measures. This will inform us to what extent the clusters extracted by K-means and FCM algorithms correspond (or not) to the theoretically labeled subtypes. The class-entropy for a given cluster k is calculated as $E_k = -\sum_{j=1}^{C} p_{kj} ln p_{kj}$, where j is one of the $C = 7$ class C GPCR subtypes and $p_{kj} = m_{kj}/m_k$, where, in turn, m_k is the number of sequences in cluster k and m_{kj} is the number of subtype j sequences in cluster k.

For FCM, Figure 1 and Table 1 show that, for the AAC data transformation, almost none of the defined clusters show clear class (subtype) specificity. Only in cluster 1, the first subtype (mG) of GPCR achieves a specificity that is close to 60%, but even in this case, the third subtype (GB) reaches a non-negligible 30%. Several clusters show common specificity profiles: for instance, clusters 1 and 3 are predominantly a mixture of mG and GB, which means that they might truly be a single cluster with some substructure. Cluster 4 is a very mixed combination of Ph and VN, but clusters 2, 6 and 7 seem to be variations of this combination, again suggesting one main cluster with further substructure and important levels of overlapping. The ACC and Digram transformations (Figures 2 and 3, and, again, Table 1), instead, manage to separate some of these clusters to become more subtype-specific. mG and GB are now more clearly discriminated (clusters 1 plus 5 and cluster 3, in turn) with the rest of subtypes showing clear overlapping in some clusters but also high specificity in others (for instance, Ph in ACC cluster 6 and Digram in cluster 7). In any case, the more complex transformations (ACC and Digram) seem to make the FCM clustering model more class C GPCR subtype-specific.

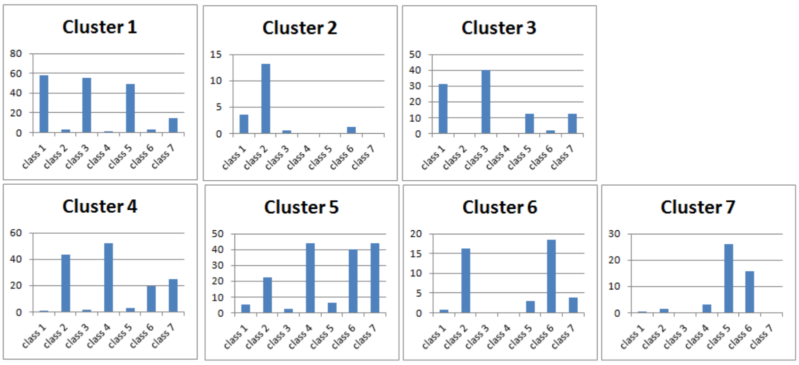

Fig. 1. Class specificity bar chart (with percentage values) for each FCM cluster of class C GPCR data set with the AAC transformation. Classes 1 to 7 are, in turn, mG, CS, GB, VN, Ph, Od and Ta.

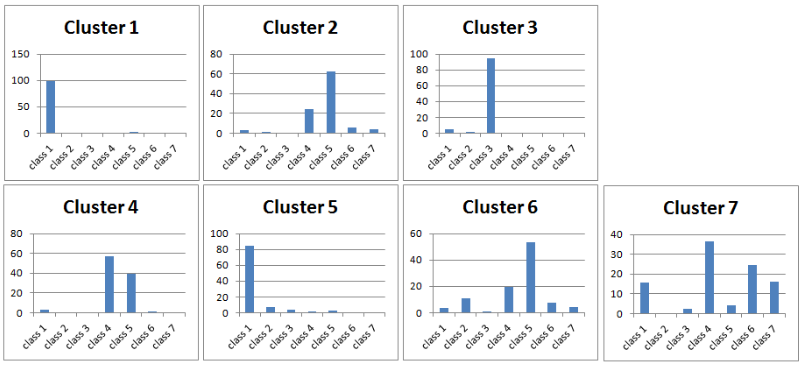

Fig. 2. As Figure 1, for the ACC transformation.

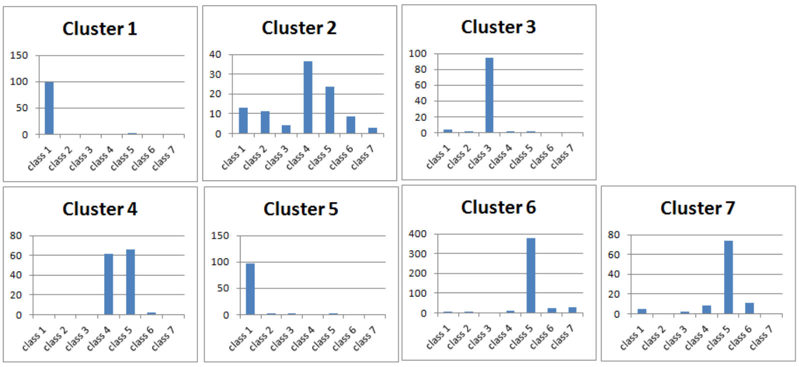

Fig. 3. As Figure 1, for the Digram transformation.

Table 1. Number of GPCR sequences-per-cluster ($\sharp$) and cluster-specific (E_k) and total entropies for the FCM clustering of class C GPCR data with the three transformations.

	AAC		ACC		Digram	
	$\sharp$	E_k	$\sharp$	E_k	$\sharp$	E_k
Cluster 1	245	1.45	107	0.13	112	0.12
Cluster 2	239	2.16	207	1.52	374	2.39
Cluster 3	200	1.34	202	0.33	200	0.37
Cluster 4	193	1.30	237	1.17	277	1.09
Cluster 5	67	1.97	199	0.89	179	0.26
Cluster 6	263	2.15	279	1.99	205	2.02
Cluster 7	303	1.95	279	2.20	163	1.28
Total Entropy	**1.77**		**1.34**		**1.29**	

The results of the K-means algorithm for the seven subtypes of class C GPCRs, for which, for the sake of brevity, we only report the entropy results in Table 2, are consistent with those of FCM. Again, almost none of the defined clusters show clear class (subtype) specificity with AAC data transformation. The ACC and Digram transformations, instead, manage to separate some of these clusters to become more subtype-specific. The similarity between the two algorithms is that mG and GB are more clearly discriminated than the rest of subtypes. Moreover, in the ACC transformation, Ph receptors can also be discriminated from the rest subtypes due to their high specificity in cluster 2. The remaining subtypes show clear overlapping in some of the clusters.

Table 2. Number of GPCR sequences-per-cluster ($\sharp$), together with cluster-specific (E_k) and total entropies for the K-Means clustering of class C GPCR data with the three transformations.

	AAC		ACC		Digram	
	$\sharp$	E_k	$\sharp$	E_k	$\sharp$	E_k
Cluster 1	270	1.65	260	0.26	222	0.10
Cluster 2	406	2.17	136	0.72	398	1.47
Cluster 3	196	1.33	150	0.23	121	0
Cluster 4	189	1.24	188	1.07	284	1.10
Cluster 5	54	1.34	165	1.57	184	1.90
Cluster 6	56	1.74	379	1.79	197	1.95
Cluster 7	339	2.03	232	1.97	104	1.63
Total Entropy	**1.77**		**1.19**		**1.39**	

Comparing the results of both algorithms in terms of the total entropy measure, conclusions are not clear-cut. ACC and Digram show a clear advantage both in FCM and K-Means, but neither shows a clear advantage over the other.

We now move to the clustering stability analyses results, based on random algorithm initializations and varying number of clusters, using the SeCo maps. For each one of the transformed sets, three SeCo maps were created using:

- The K-means objective function and the standard Cramér's V index.
- The FCM objective function and the standard Cramér's V index.
- The FCM objective function and the novel weighted Cramér's V index proposed.

As previously mentioned, a threshold for the $\triangle SSQ$ values to select the 10% top values for each value of K is expected to allow the degeneracy of similar SSQ values to be resolved. The FCM 10% top results, as reported in Figs. 4 to 6, are very parsimonious (much more so than the complete ones, not reported here), revealing a high concentration of stability results around just a handful of median Cramérs V index values, in comparison with the still wide spread of K-Means. These results are also very consistent over data transformations. For K-Means, this effect does not necessarily increase as K increases for any of the data transformations. For FCM, though, an increase in spread as K increases is revealed. Overall, this indicates that FCM is much more resilient than its crisp K-Means counterpart to the variability introduced by random initializations.

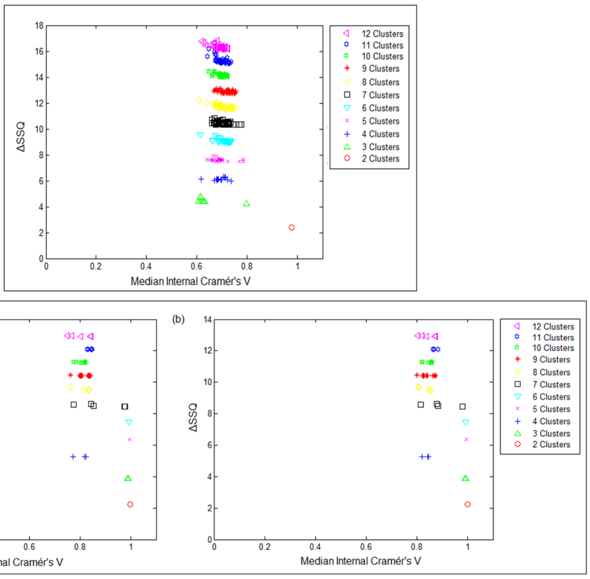

Fig. 4. Separation-Concordance maps for the AAC data set, including the 10% best results. Top: results for K-Means; bottom: results for FCM, a) with standard Cramér's V index; b) with proposed weighted Cramér's V index.

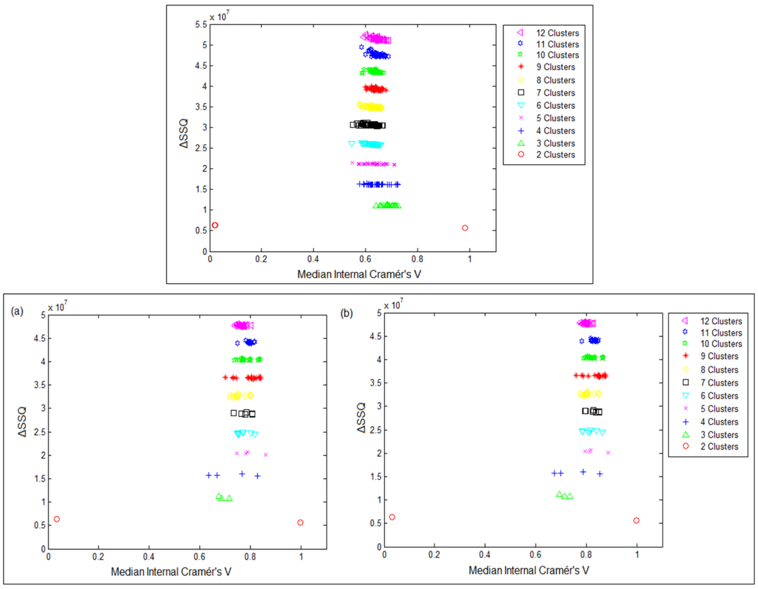

Fig. 5. Separation-Concordance maps for the ACC data set. Layout as in previous figure.

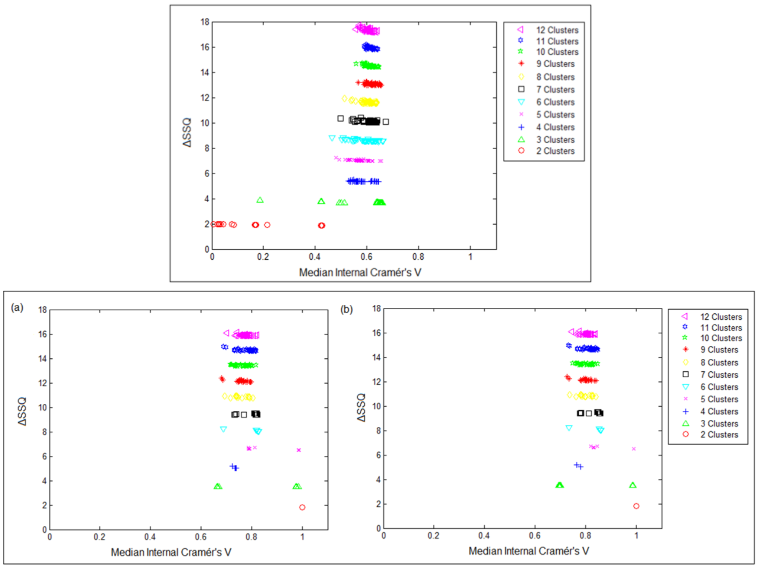

Fig. 6. Separation-Concordance maps for the *digram* data set. Layout as in previous figure.

Moreover, the stability results as measured by the standard Cramérs V index and the proposed weighted Cramérs V index are again very similar for all data transformations (even better for the latter, providing further support for the proposed method, which is a more faithful account of the true belief of the algorithm regarding cluster membership).

Also, the restricted 10% SeCo maps offer some guidance to make a decision about the most adequate value of K, as supported by the data. For FCM, a solution beyond 7 clusters is clearly not supported by the AAC transformation, as maximum stability suddenly decreases at the same point the cluster model becomes more unstable (with more spread values). Note that this is consistent with the "natural" description of subtypes for this data set. A similar conclusion, though, is not supported for ACC and only partially for Digram, whose low-K solutions are clearly polarized. In any case, these results are hardly conclusive, which means that SeCo maps have limited applicability for the choice of K in highly overlapping data sets such as those analyzed in this study.

4 Conclusions

Crisp clustering provides a simplified partition of the observed data as a description of its structure. K-Means, the most commonly used algorithm of this type, is known to be strongly dependent on initialization. Furthermore, recent studies suggest that different K-Means solutions with apparently similar error may in fact be quite dissimilar. In this context, it is recommended to base the choice of solution on a combination of error and stability criteria.

Separation-Concordance maps provide such combined criterion. They were originally developed for K-Means and, in this study, we have extended them to FCM by defining weighted contingency tables and a weighted Cramér's V index that could also be used in other fuzzy and probabilistic techniques in which cluster assignment is not of a *crisp* nature any longer.

We have experimented with this approach in a problem concerning the clustering of class C GPCRs, which have a pre-defined subtype structure. This sub-typology is known to have a highly overlapping structure from a clustering viewpoint, which has been confirmed in our experiments. The SeCo maps have revealed the FCM algorithm to yield much more stable results that K-Means under multiple random initializations, but they have also been shown to provide limited guidance for the choice of the K and C parameters, due to the highly overlapping nature of the data. In any case, the proposed weighted Cramér's V index provided consistent with and often better results than the standard one in our experiments. This is encouraging, given that the modified index is meant to reflect the nature of the clustering results more faithfully, something that might have revealed lower stabilities.

References

1. M. Rask-Andersen and M. Sällman-Almén and H.B. Schiöth, Trends in the exploitation of novel drug targets, Nature Reviews Drug Discovery, 10, 579–590, 2011.

2. V. Katritch and V. Cherezov and R.C. Stevens, Structure-function of the G Protein-Coupled Receptor superfamily, Annual Review of Pharmacology and Toxicology, 53, 531–556, 2013.
3. K. Hollenstein *et al.* Structure of class B GPCR corticotropin-releasing factor receptor 1, Nature, 499, 438–443, 2013.
4. F.Y. Siu *et al.* Structure of the human glucagon class B G-protein-coupled receptor, Nature, 499, 444–449, 2013.
5. H. Wu, *et al.* Structure of a class C GPCR metabotropic glutamate receptor 1 bound to an allosteric modulator, Science, 344(6179), 58–64, 2014.
6. A.S. Doré, *et al.* Structure of class C GPCR metabotropic glutamate receptor 5 transmembrane domain. Nature, 551, 557562, 2014.
7. J. Kniazeff and L. Prézeau and P. Rondard and J.P. Pin and C. Goudet, Dimers and beyond: The functional puzzles of class C GPCRs, Pharmacology & Therapeutics, 130, 9–25, 2011.
8. R. Karchin and K. Karplus and D. Haussler, Classifying G-protein coupled receptors with support vector machines, Bioinformatics, 18, 147–159, 2002.
9. B. Liu and X. Wang and Q. Chen and Q. Dong and X. Lan, Using amino acid physicochemical distance transformation for fast protein remote homology detection, PLoS ONE, 7, e46633, 2012.
10. M.I. Cárdenas, A. Vellido, C. König, R. Alquézar and J. Giraldo, Exploratory visualization of misclassified GPCRs from their transformed unaligned sequences using manifold learning techniques, In Procs. of the IWBBIO 2014, 623–630, 2014.
11. R. Cruz-Barbosa, A. Vellido, and J. Giraldo. The influence of alignment-free sequence representations on the semi-supervised classification of Class C G Protein-Coupled Receptors. Medical & Biological Engineering & Computing. In press, doi: 10.1007/s11517-014-1218-y
12. A.K. Jain, Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651-666, 2010.
13. J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Kluwer Academic Pub. 1981.
14. P.J.G. Lisboa, T.A. Etchells, I.H. Jarman and S.J. Chambers, Finding reproducible cluster partitions for the k-means algorithm. BMC bioinformatics, 14(S1), S8, 2013
15. A. Gutcaits, *et al.* Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. Protein Science, 11(4):795-805, 2002
16. M. Sandberg, L. Eriksson, J. Jonsson, M. Sjöström, and S. Wold. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. Journal of Medicinal Chemistry, 41(14):2481-2491, 1998
17. C. Caragea, A. Silvescu, and P. Mitra, Protein sequence classification using feature hashing, Proteome Science, 10.Suppl, S14, 2012
18. A. Ben-Hur, A. Elisseeff, and I. Guyon, A stability based method for discovering structure in clustered data. In Pacific Symposium on Biocomputing, Vol. 7, pp. 6-17, 2001
19. B. Vroling, *et al.* GPCRDB: information system for G protein-coupled receptors, Nucleic Acids Research, 39(suppl 1):D309-D319, 2011.
20. C. König, R. Cruz-Barbosa, R. Alquézar and A. Vellido, SVM-based classification of class C GPCRs from alignment-free physicochemical transformations of their sequences. Lecture Notes in Computer Science 8158, pages 336–343, Springer, 2013.