Pattern Recognition Letters 74 (2016) 24–29

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Using a novel clumpiness measure to unite data with metadata: Finding common sequence patterns in immune receptor germline V genes☆

Gregory W. Schwartz[a], Ali Shokoufandeh[b], Santiago Ontañón[b], Uri Hershberg[a,c,*]

[a] Department of Biomedical Engineering, Science & Health Systems, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA
[b] Department of Computer Science, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA
[c] Department of Microbiology and Immunology, Drexel University College of Medicine, 2900 W. Queen Lane, Philadelphia, PA 19129, USA

ARTICLE INFO

Article history:
Received 20 June 2015
Available online 8 February 2016

Keywords:
Hierarchical clustering
Aggregation
Tree analysis
Immune receptor repertoire
Adaptive immunity
Multiscale analysis

ABSTRACT

When finding relationships in biological systems, we often describe hierarchies based on one facet of the data. However, when using this hierarchy to elucidate relationships between metadata, the distribution of metadata labels within the hierarchy may exhibit different levels of aggregation—uniform, random, or clumped. As of now, there exists no measure for finding the level of aggregation, or "clumpiness", between labels distributed among the leaves of a hierarchical container. We propose a clumpiness measure to aid in the quantification of relationships between metadata. We validated our measure with random trees and found that the measure is resistant to changes in the tree size, label size, and the number of types of labels, compared to the closest alternative measures. We used our clumpiness measure to quantify the relationships between light and heavy chains in human and mouse B cell and T cell receptor V genes based on their motifs. We found that the B cell heavy chains were the most aggregated while the T cell chains were the least aggregated and that the IGL chain was clumped the most with the T cell chains out of all of the B cell chains.

1. Introduction

Biological systems are often described through hierarchical relationships of different labels. Due to the complex nature of biological systems, we often describe these hierarchies using only a subset of the available information about each element in the dendrogram. For instance, we can create a phylogeny of different species based on their genes while at the same time retaining other metadata, or labels, about their behavior, survival, and phenotypes. However, unlike the gene data, these metadata labels may be distributed randomly, uniformly, or clumped throughout the hierarchical structure. In this paper, we present a novel measure to quantify the extent that a hierarchical structuring of data describes a relationship of aggregation, or "clumpiness", between the metadata labels with which its components are categorized. In this fashion, we can unite the two levels of the structure—the information from the data and the categorical information from the metadata.

Let us consider the adaptive immune system as a general example of a multi-scale biological system. This system is comprised of several repertoires of immune cells with individual receptors of unique specificity for different antigens in the environment. In order to cover a wide range of antigens, the body generates a vast and diverse pool of differently responding cells called the immune repertoire of the body. Under specific conditions, these antigens can trigger competitive proliferation, mutation, and death in only a subset of the cells. The successful recognition of an antigen by a cell's receptor leads to the cell dividing and producing its own lineage of cells responding to similar antigens. The resulting hierarchical structure is associated with metadata labels such as the tissue where one of the descendant cells is found, the function of that cell in the immune response (such as an effector cell or a memory cell), or its fate—death or division. Because the metadata labels are of a different scale than the data (in this scenario the pattern of mutations in a given cell), it is possible to have the labels widely dispersed in the container (here a hierarchical data structure) but be close together in small clumps as opposed to

☆ This paper has been recommended for acceptance by Qian Xiaoning.

* Corresponding author at: Department of Biomedical Engineering, Science & Health Systems, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA. Tel.: +1 215 895 1698, fax: +1 215 895 4983.
E-mail addresses: gregory.schwartz@drexel.edu (G.W. Schwartz), uri.hershberg@drexel.edu (U. Hershberg).

http://dx.doi.org/10.1016/j.patrec.2016.01.011
0167-8655/

being randomly or uniformly distributed. These scenarios appear throughout the biological domain.

Another hierarchical container of data we may use, in order to capture the different possible "behaviors" of the cells, is to cluster the cells by their common gene expression patterns [1]. In this case, the labels describe a common progenitor (ancestor) cell or varying levels of mutation in its DNA. Finally, we will look specifically in this paper at the hierarchical clustering of sequence fragments. Both in our example and in other biological examples, it is commonly considered that motifs can be indicative of binding capability and interaction potential. In this case, we hypothesize that a group of cells with similar behaviors are motivated by a subset of common motifs with related structures. As binding of the receptor and survival of the cell depend on the receptor structure, in order to look at the relationships between the different parts of the receptors we need to account for the relationship between sequence fragments. As such, we must find the relationship (container) between each region of the receptor (data) before quantifying the overall distribution and degree of aggregation of chains (label).

As shown in these examples, although the data is clustered together as the result of the tree, we can ask additional questions about the relationship between the metadata labels within the tree. More specifically, we would like to quantify the degree of aggregation, or "clumpiness", between the labels by using the structure of the tree generated by the pairwise relationship of the data. While there is a wide range of metrics to measure aggregation, they are focused on the spatial distribution in two dimensions [2–9]. As of now, there exists no measure for the quantification of aggregation in a distribution within hierarchical trees. Furthermore, previous studies attempting to find patterns between metadata in hierarchical structures are based on grouping similar sections of the container rather than finding the impact of dispersion on the metadata and are heavily focused on visualization [10–13]. In this paper, we will demonstrate the power of such an analysis by focusing on the last example, where we can find the relationship between immune receptors by their sequence fragments.

In order to look at the distribution of labels, we need new tools. We propose our clumpiness measure as a way to measure the degree of aggregation between labels in a hierarchical container. Our measure is robust to the container size, data size, and label size, and thus is scale invariant. In addition, our measure is generalizable to more than two labels and is efficiently computable and maintainable. In this paper, we will (1) describe the measure, (2) show the generalization, (3) demonstrate the use of the measure to find the relationship between receptor chains, and (4) quantify the response of the measure to noise and sizes.

2. Notations and definitions

Suppose we have a rooted binary tree with a set of vertices V. Let us now call $I \subseteq V$ the set of non-leaf and non-root vertices of the tree, and $T \subseteq V$ all of the leaf vertices whose parent is in I (thus, this includes all the leaf vertices, except the leaves that are children of the root, since the root is not in I). Now, let us assume $M \subseteq T$ to be the subset of leaves of interest, our "relevant" leaves, and $L = \{L_1, L_2, \ldots, L_n\}$ to be a partition of M (i.e., $M = \bigcup_{i=1}^{n} L_n$). This partition can represent, for example, a set of labels that we care about in our application domain. We call these labels "relevant" as they contain our relevant leaves. We can now transform the data from our domain into a hierarchical container.

We specifically focus on the domain of immunology in this study. The B and T cells are white blood cells with cell surface receptors, the B cell receptor (BCR) and T cell receptor (TCR) respectively, that bind to antigen which can invoke an immune response. These receptors are quite diverse and each B and T cell express just one type of this receptor. The BCR is composed of a heavy chain (IGH) and a light chain (either IGK or IGL), while the analagous chains on the TCR are the β chain (TRB) and the α chain (TRA). As we want to compose our hierarchical container from structural units, we use subregions of these receptor genes in our clustering.

These subregions, we call "protein fragments", are 20 amino acid long sequences taken from an overlapping sliding window across an amino acid receptor sequence. Then our hierarchical clustering generates clusters that are each a group of protein fragments with similar sequences (further explained in Section 5.1). The leaves in the hierarchical container represent these clusters, where each parent contains the union of the children's protein fragments. In this way we have completed the transformation of the data into a hierarchical container of relationships.

3. Clumpiness measure

3.1. Definition

The clumpiness of the set of leaves M when partitioned according to L in a k-ary tree is defined as

$$C(L) = \frac{1}{n}\left(\prod_{i=1}^{n} \frac{x}{y_i}\right)^{1/n} \tag{1}$$

That is, the geometric mean of x weighted by the frequency of each label, y_i. The result is set between 0 and approximately 1 by normalizing by the total number of labels, n. The numerator x is intuitively the weighted number of viable vertices in I weighted by y_i, resulting in

$$x = \frac{1}{|I|}\sum_{v \in I} \delta(v) w(v) \tag{2}$$

$$y_i = \frac{|L_i|}{|T|} \tag{3}$$

We say that a non-root vertex v is "viable" if $\delta(v) = 1$, meaning that v has at least one vertex of each label in its descendant leaves. So,

$$\delta(v) = \begin{cases} 0 & : \bigvee_{i=1}^{n} |D(v) \cap L_i| = 0 \\ 1 & : \text{otherwise} \end{cases} \tag{4}$$

where $D(v)$ is the set of descendant leaves of vertex v contained in M, our relevant leaves. We then weigh the vertex if it is viable by the number of vertices of each relevant label and how far away they are from the vertex in question

$$w(v) = \sum_{i \in D(v)} \left(\prod_{j \in E(v,i)} \frac{1}{c(j)} \right), \tag{5}$$

where $E(v, i)$ is the set of vertices on the shortest path from (and including) v to (but not including) the relevant leaf i and $c(j)$ is the number of children of j. We weigh by the number of children as we want the maximum value of our vertex of interest to be 1, so we keep dividing the values of the descendant vertices based on branching.

If we want to find the clumpiness of a label L_i with itself we need to change our approach: the more clumpy L_i is with other labels, by definition the less clumpy L_i is with itself. Using this property, we can then have L contain two sets—those leaves in L_i and all other leaves. Then the clumpiness of L_i with itself becomes $1 - C(L)$. For the sake of simplicity, we will focus on the case of a rooted full binary tree containing 2 labels.