SUPPLEMENTAL INFORMATION

Generating Information Rich High-Throughput Experimental Materials Genomes using Functional Clustering via Multi-Tree Genetic Programming and Information Theory

Santosh K. Suram, Joel A. Haber, Jian Jin and John M. Gregoire

In this article, we propose a new Cauchy-Schwarz divergence function that is invariant to number of clusters. To illustrate the applicability of our approach, we assign random memberships for various number of clusters (2 to 8) to 486 compositions distributed at a 3.33 at.% interval in a ternary library; and plot the cross cluster information potential, self-information potential and their ratio using the Cauchy-Schwarz divergence function D_{cs} proposed by Boric et al.[1] (Eq. 1) and the Cauchy-Schwarz divergence function proposed in our article (Eq. 2) in Fig.1 and 2 respectively.

$$D_{cs}(p_1, p_2, \ldots . p_c) \approx -\ln \frac{\frac{1}{2}\sum_{i,j=1}^{n}(1 - {}^{i}\mathbf{m}^{\mathrm{T}}\,{}^{j}\mathbf{m}) G_{ij,2\sigma^2}}{\sqrt{\prod_{k=1}^{c}\sum_{i,j=1}^{n} {}^{i}m_k\,{}^{j}m_k G_{ij,2\sigma^2}}} \qquad [1]$$

$$D_{cs}(p_1, p_2, \ldots . p_c) \approx -\ln \frac{\sqrt{\left(\sum_{i,j=1}^{n}(1 - {}^{i}\mathbf{m}^{\mathrm{T}}\,{}^{j}\mathbf{m}) G_{ij,2\sigma^2}\right)\left(\frac{c}{c-1}\right)}}{c\left(\prod_{k=1}^{c}\sum_{i,j=1}^{n} {}^{i}m_k\,{}^{j}m_k G_{ij,2\sigma^2}\right)^{\frac{1}{2c}}} \qquad [2]$$

From Fig. 1 we see that the self-information potential increases as a power of number of clusters (c); whereas the cross information potential increases very slowly as a function of c. Thus, the self-information potential dominates the Cauchy-Schwarz divergence function as the number of clusters increases.

In Fig. 2; we show that the Cauchy-Schwarz divergence function proposed in our approach results in cross-information and self-information potentials that are invariant to number of clusters (c). Further, the ratio of cross-information potential to self-information potential is approximately 1 resulting in approximately zero Cauchy-Schwarz divergence; as expected for compositions with random memberships.

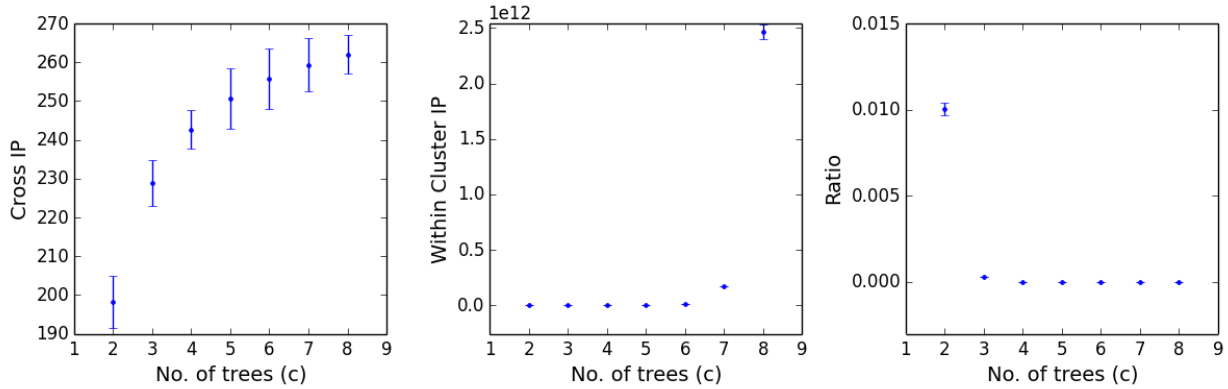


Fig1. Plot of cross information potential, within cluster (self) information potential and their ratio using the Cauchy-Schwarz divergence function suggested by Boric et al[1]. Error bars are based on standard deviations obtained by assigning random memberships 10 times.

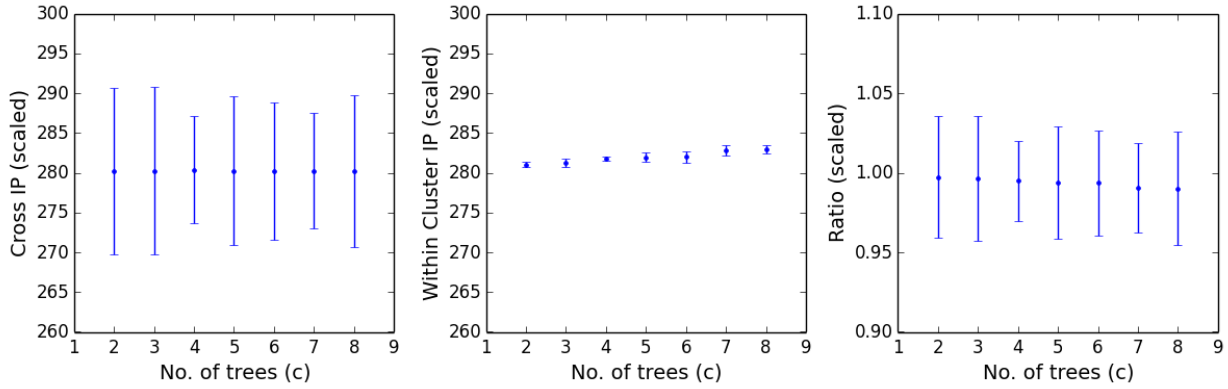
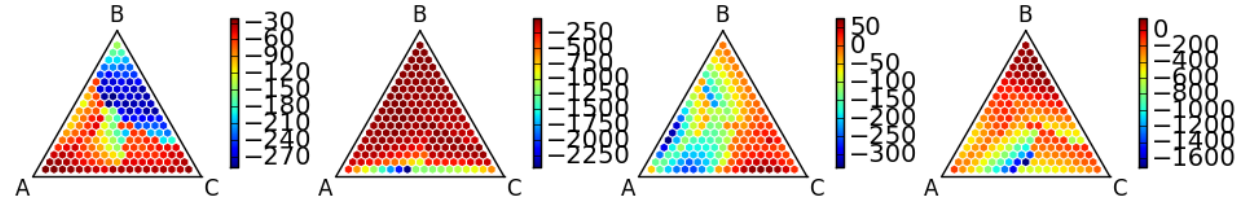


Fig 2. Plot of cross information potential, within cluster (self) information potential and their ratio using the modified Cauchy-Schwarz divergence function suggested in our article. Error bars are based on standard deviations obtained by assigning random memberships 10 different times.



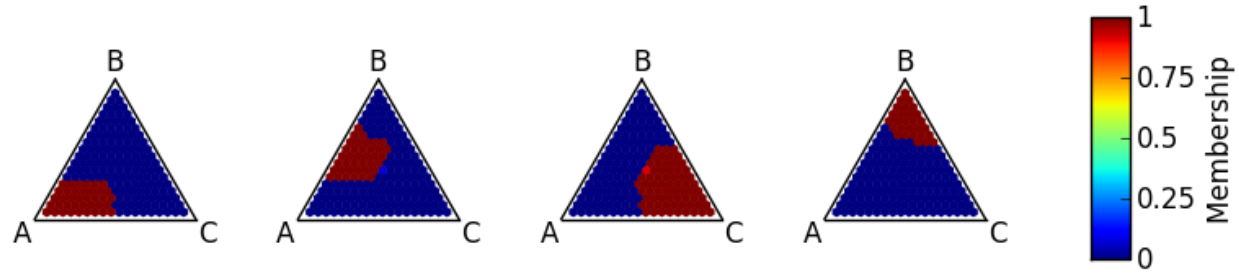


Figure 3. top: False color maps of outputs of genetic programming trees for four different trees. bottom: Corresponding membership functions. The clustering was carried out using sigmoid transformation between genetic programming outputs and membership values as suggested by Boric et al.[1] instead of the linear scaling approach suggested in our approach. Loss of information during the sigmoid transformation is easily observed.

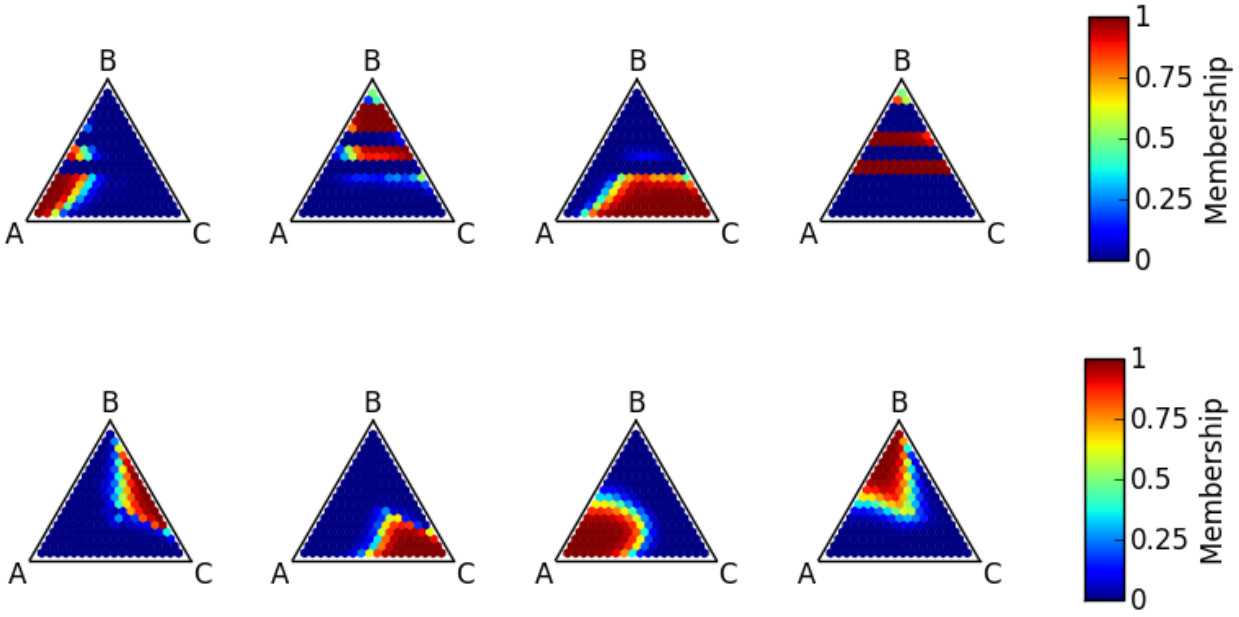


Fig. 4 Result of our clustering algorithm using a Gaussian Parzen kernel width of (top) σ=0.1 and (bottom) σ=0.3. σ=0.1 results in very noisy clustering thus indicating that 0.1 is a much smaller than required kernel width. Whereas, σ=0.3 results in clusters that misrepresent several property fields indicating that 0.3 is a larger than required kernel width.

References

(1) Boric, N.; Estévez, P. A. Genetic Programming-Based Clustering Using an Information Theoretic Fitness Measure. **2007**, 31–38.