Inferring Key Biological Tasks in Acute Myeloid Leukemia Using Multi Objective Optimization

David Alber[a], Lukas Alber[a]

[a] *Department of Physics, Stockholm University, 106 91 Stockholm, Sweden*

Abstract

In nature, biological systems subjects to diverse tasks often face a trade off; specialization on one phenotype can not be optimal at all tasks. Using Archetypal Analysis (AA), we analyze acute myeloid leukemia (AML)- tumor samples and infer key biological tasks cells can specialize, in order to outperform competitors and dominate its niche. AA is an unsupervised learning method, which allows data to be represented as convex combinations of external elements of the data set. As opposed to many other cluster methods, optimization with respect to multiple extrema allows for consideration of mutually exclusive features and can therefore account well for evolutionary trade- off scenarios. Since certain features close to the archetypes are maximally enriched, they can be linked to respective tasks. We demonstrate that AML- tumors are well described by a triangle, suggesting three main traits to specialize in.

Keywords: Pareto Optimization, Multi Objective Optimization, Biological Trade- Offs, Archetype Analysis, Principal Convex Hull Analysis, Tumor Biology, Clustering

1. Introduction

Multi objective- (or pareto) optimization is a machine learning method, that allows to optimize with respect to multiple extrema. Typically, machine learning algorithms are designed in a way such that, for a model with certain complexity Ω, they are trying to minimize the loss function $\mathcal{L}$, often the mean squared error. In clustering, unsupervised settings, the task is often to maximize the similarity of members within the same group while minimizing the similarity of elements in different clusters.

A multi objective optimization approach allows to optimize with respect to multiple features. I.e. instead of choosing a model Ω with a certain complexity, and minimizing the loss function $\mathcal{L}$, pareto optimization allows to optimize with respect to both, model complexity and loss, or multiple features [1]:

$$min(\Omega, \mathcal{L}) \tag{1}$$

Especially in a biological framework, where specimen often face a trade off between mutually exclusive tasks, similar to the complexity- loss trade off, this approach is very powerful. Living things can specialize in different tasks in order to survive, analyzing them with a multi objective optimization approach has been introduced in the framework of Archetypal Analysis (AA) by [2] and is current subject to research as in [3] and [4]. Here one faces the unsupervised learning task of trying to best characterize the data with respect to multiple extrema, Archetypes. That is, what are the distinct tasks the subject under consideration can specialize in.

The high dimensional biological data is described as a point cloud, where each point, tumor sample, is determined by the measured gene expressions of the respective sample. AA allows to enclosed those samples by a polytope. The vertices (or archetypes) of the polytope are extreme points and give rise to the identification of the distinct tasks the samples under consideration have to achieve in order to be biologically *successful* [5]. As all the points within the polytope and on the surface, pareto front, represent convex combinations of the archetypes also those specimen are biologically *successful*, since they specialize in a mixture of the distinct tasks. Only the points outside of the polytope are said to have bad performance in all tasks and therefore less *fitness* [3].

Even though data pre- processing is often disregarded in other reports, we find it essential to mention and discuss it in section 2. There we describe and motivate our applied transformations and our dimension reduction method, i.e. principle component analysis (PCA), which rotates the data into the components that explain the most variance, allowing us to go from high- to lower dimensional space Then, after briefly mentioning K-means clustering as an intuitive 'first try' approach in section 3, we guide the reader through a three step process of AA in section 4. First we use a Principal Convex Hull Analysis (PCHA) algorithm to find the coordinates of the vertices, i.e. archetypes. Second we back-transform the identified archetype coordinates from PCA space to the full gene expression space and map them to Gene Onthology- (GO) space using the

Email addresses: alber.p.david@gmail.com (David Alber), lukas.alber@fysik.su.se (Lukas Alber)

Molecular Signature Database (MSigDB) [6]. Third, after having found the coordinates of the vertices, in the GO-space we aim to find the corresponding biological tasks by different measurements. That is, we analyze the enriched features. In the last section 5 we aim to give a biological interpretation for the data under investigation.

2. Data pre-processing

The way a given data set is pre-processed heavily influences the way it can later be interpreted. The data set under consideration consists of 451 samples of blood cancer patients (AML- tumors). Each sample contains information on how well 22843 genes are expressed. These characteristics vary widely in terms of its magnitude of expression. In order to make the variation of magnitudes comparable we apply two, in biology commonly used transformations, to the original data:

First we add an offset of +1 and then we apply an element wise natural logarithmic transformation to the data. Since a high gene expression does not automatically mean high importance, this allows to scale all measured data to comparable size.

Second, for every characteristic we subtract the mean over all samples. That is, for every of the 22843 different gene expressions we are interested in the amplification with respect to the mean value of this expression.

As a consequence to this non linear transformation highly expressed gene characteristics scale down to smaller values. The structure of the data is scaled to avoid extremities and the data points are generally more packed after the transformation is applied. Further more does the *squeezing* of the logarithmic scaling quench outliers, visible in the *raw* data depiction. This transformation makes it harder to find distinct structural patterns in the data, since unsupervised learning tasks perform better on well separated data. Classification is easier if the data shows separable features. However, for the result to be interpretable in a biological sense, the transformation is an absolute necessity. High expression of some genes does not automatically mean that these characteristics are more representative for a particular specialisation. It could also be an artefact, since maybe for some tasks more active genes are involved than for another, without this task itself being more important. The log- transformation together with the relative deviation from the respective mean give therefore more reliable information on the importance of the tasks for which the genes correspond to.

In order to reduce the dimensionality of the data we perform a PCA to the now already log- transformed and mean- deduced data. This procedure is also known to be a good practice to deal with noise in the data, since the PCA decomposes the data in the components that carry the most information. This allows to disregard dimensions that carry little information and are associated to noise. Figure 1 gives a information on how many PCA- components are a reasonable choice for this decomposition. According to fig. 1 we conclude that 40 PCA-components, which account for more than 70% of the information, are a good choice for further analysis.

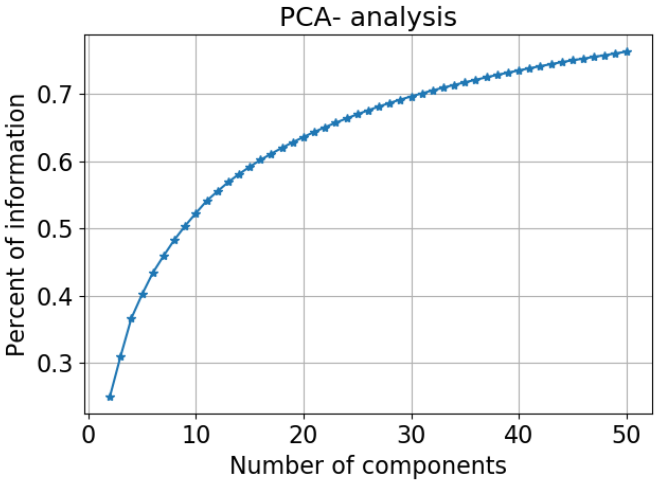

Figure 1: Cumulative percent of explained variance as function of PCA- components for the given data set of originally 451 samples and 22843 characteristics.

3. K-means Clustering

An intuitive and commonly used approach for unsupervised learning tasks is to use the structure of the data to classify the samples to a set of clusters, i.e. the data is categorized by its spacial structure. A general challenge for unsupervised learning tasks, where no prior knowledge of the number of clusters is provided, is to decide the best number of clusters. The number of clusters should be large enough to explain the structure in the data but not be too large, since redundancy is tried to be avoided. As a quantitative measurement for the goodness of the chosen amount of clusters we use the *Silhouette number s*. According to the distances from one sample point to points, assigned to the same cluster and points assigned to a different cluster, the *Silhouette number* yields a score from $s \in [-1, 1]$. A *Silhouette* score close to +1 indicates that the sample is far away from the neighbouring clusters and hence well assigned. Negative *Silhouette* scores indicate, that the samples in question are far away, by means of euclidean distance, from the cluster to which it is assigned and close to a different cluster. More precisely the *Silhouette number* $s(i)$ of sample i is defined as follows [7]:

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ \frac{a(i)}{b(i)} - 1, & \text{if } a(i) > b(i) \end{cases} \tag{2}$$

In eq. (2) $a(i)$ denotes the mean distance between data point i and all the data points assigned to the same cluster C_i and $b(i)$ the data point i and all the data points assigned

Figure 2: Analysis of the best number of clusters based on the average *Silhouette number*. The data used for the analysis are 451 blood cancer samples, PCA reduced from 22843 to 40 characteristics which corresponds to more than 70% of the explained information. A peak of the average *Silhouette number* can be identified for 18 clusters.

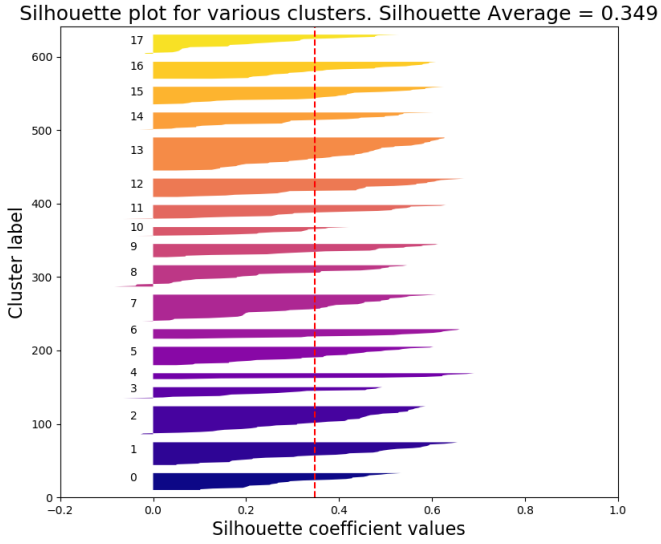

Figure 3: *Silhouette number* of every sample associated to 18 different clusters, which are represented by the different colors. The dashed red line represents the *Silhouette* average of all samples for 18 clusters.

to the nearest neighbouring cluster C_k i.e.:

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, j \neq i} d(i,j) \tag{3}$$

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$$

The K-means clustering algorithm is used for the data set, which has been PCA reduced from originally 22843 characteristics to 40 PCA components, explaining $\approx 72\%$ of the variation in the data, see fig. 1. We compare the average *Silhouette* score for different amount of clusters in fig. 2. The low average scores indicate that the assignment of the data points to the clusters is not suitable. We are looking for a plateau in of the *Silhouette number* where a rather small number of clusters is chosen. Since one can see a peak in fig. 2 for 18 clusters, we show a more detailed depiction of this case in fig. 3 and 4.

The thickness of the bars in fig. 3 indicate the amount of data points assigned to the different clusters, indicated by the color. One can observe that the *Silhouette* scores are, besides a small number of exceptions, always positive, but still not close to 1, indicating that the assignment of clusters is not obvious. Indeed, the graphical 2D- projection of the sample data see fig. 4, shows that the structure of the data does not support a clear and intuitive spacial separation of the data.

Since the graphical depiction of the sample data does not support the idea of a clear separation to a specific number of distinct clusters and the average *Silhouette* scores for various numbers of clusters is low, we conclude that finding a structural pattern for the given data is not ideal with K-means clustering. More over, we want to describe our data with as little as possible representative points. Therefore, we use AA to find distinct, exclusive features in the samples.

4. Archetype Analysis

Opposed to K-means cluster methods, where typical features, cluster centers, of the data set under investigation are found, Archetype Analysis (AA), proposed by [2], seeks external points in the multi dimensional data. The archetypes found can then represent the data via convex combinations.

As mentioned above, we are specifically interested in finding distinct features of the AML- data set under investigation. That is rather than finding typical representations in the gene expression space we want to find which distinct features can best represent our data set. Those extremal points (Archetypes), i.e. vertex points of the convex hull enclosing the data, can be biologically interpreted and linked to different key tasks, which the tumor can specialize, in order to be *successful*. As some of those features might be mutually exclusive, we want to take advantage of a multi objective optimization approach. Similar to the complexity- variance- trade off mentioned in the introduction, also biological specimen often face a trade off. That is, some cancer tumor might specialize in the mutually exclusive features of hiding from immune cells or in growth, or a convex combination, but will not be able to be expert in both. The archetypes can be related to those tasks. Using an AA approach to describe the data set as a polytope and interpreting maximally enriched features as a biological task, has already shown to be successful when dealing with high dimensional biological data and more specifically, for the classification of human breast cancer [5].

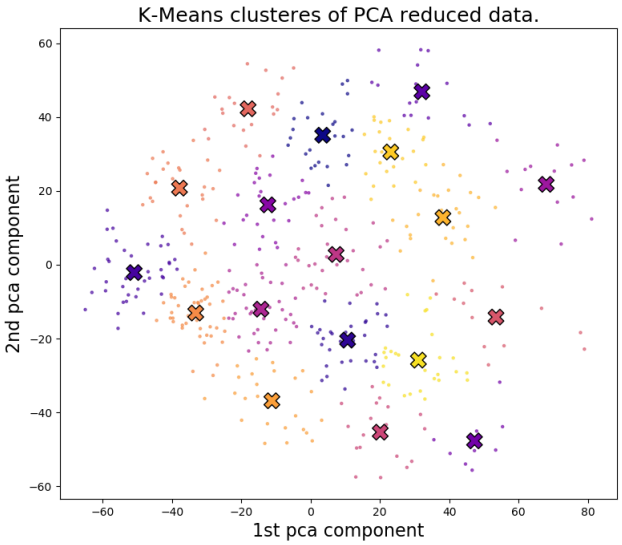

Figure 4: 2D- projection of data clusters and respective centers evaluated with a K- means algorithm. The number of clusters (18) is chosen according to the peak that can be observed in fig. 2.

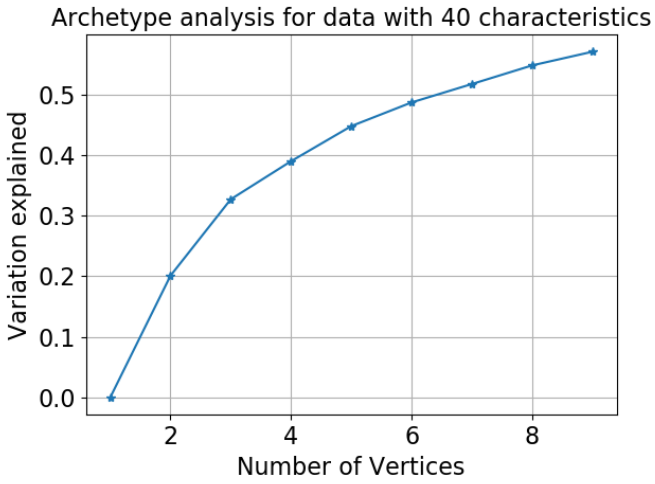

Figure 5: Explained variation with respect to different choices of vertex points used in the PCHA analysis. An elboww', smaller slope, can be observed after 3 vertex points.

4.1. Principal Convex Hull Analysis

Inspired by the work of Yuval Hart et al. [5] we also use *Principal Convex Hull Analysis* (PCHA) [8]. Similarly to K-means clustering, where the number of clusters has to be decided, the first question to clarify is, how many archetypes are expected for the given data set, i.e how many vertices does the polygon of the convex hull consist of. We want to find the ideal number of vertices for which its hull includes as many data points as possible by using as less vertices as possible. Figure 5 shows the analysis on which our choice of 3 archetypes is based on.

Since the explained variation starts to saturate for more than 3 archetypes we do not assume more. This eye measurement is known as the 'elbow test' [5]. More archetypes result in redundancy of its explained features and adding more than 3 vertices to the PCHA model contributes little to the explained variation[1].

Figure 6 gives a graphical depiction of the convex hull enclosing the given data. The polygon with 3 vertices corresponds to a triangle when projecting the data from 40 PCA-components to the first two PCA-components. Note that the PCHA was done with 40 PCA-components, which account for more than 70% of the variation. Hence the archetypes are 40 dimensional points in the PCA-reduced gene expression space.

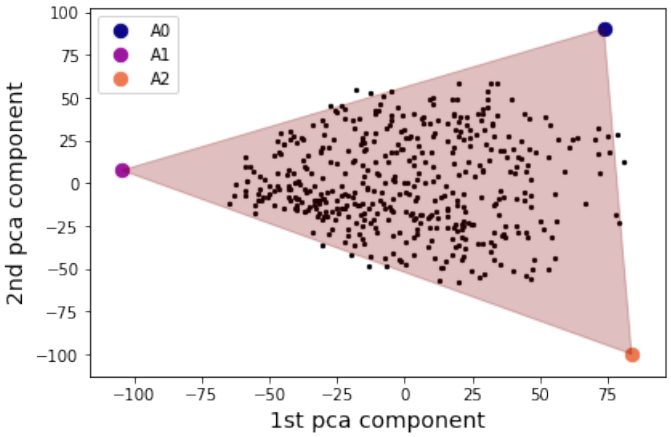

Figure 6: Projection of the high dimensional gene- expression tumor samples onto the first two principle components. The three vertex points/ archetypes highlighted correspond to distinct features and span the triangle of the pareto front, such that all points within can be explained as convex combinations of the vertices.

4.2. Gene Onthology-Transformation

Inverse transforming the coordinates of the archetypes from the 40 PCA-reduced space to the full 22843 dimensional gene space gives us information about which genes

[1] A detailed analysis of the 4 archetype case, revealing that the extra archetype is a mixture of the considered 3, has bee done.

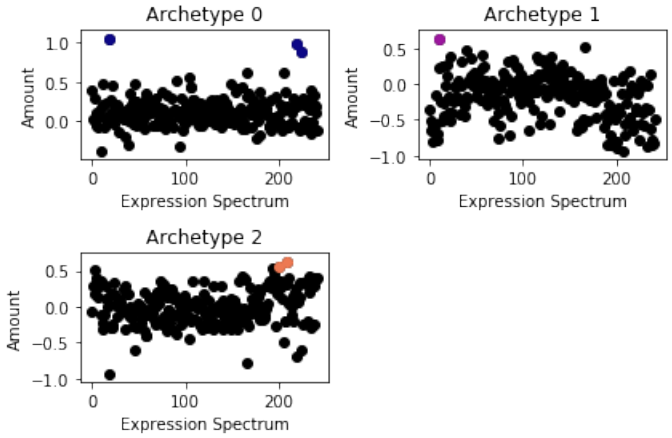

Figure 7: Visualization of the GO- expression spectrum for each archetype, showing the entire GO- space and the respective amount of expression for each dimension. Maximally expressed dimensions, colored dots, are considered to exhibit more than 85% of the global maximum, for each archetype.

represent an archetype and how strongly each gene is expressed for each archetype. However, rather than one gene being responsible for one biological task, it is often the case, that a composition of several genes is involved in a biological processes. It is therefor hard to conclude from the knowledge of gene expression to a particular biological task. For the biological interpretation another transformation is done: We trans from the coordinates from the gene expression space, as given in the raw data set, to the space of gene ontology (GO) groups, to find out which composition of genes are responsible for the execution of a particular biological process. That is, each GO- group, representing a well understood biological task, consists of multiple gene expressions. We perform a mapping from 22843, not easily interpretable gene expressions to 242, well understood GO-groups, using the Molecular Signature Database [6]. How strongly each feature in the GO-space is expressed tells us how important this task is. In order to depict this, we show the full spectrum of the entire GO- space for each archetype respectively in fig. 7, where the maximally enhanced GO- expressions are highlighted.

4.3. Enhanced GO- Expressions

We assume that the magnitude of expression of the different GO- groups gives information about which biological task each archetype is specialized on. In particular we hypothesise that a maximal expression in the GO- space can be identified as a specialisation of the corresponding archetype. Unfortunately fig. 7 shows that no archetype shows a clear, free standing maximum and a single specialisation can therefore not be identified by solely looking at the GO- expression spectrum of the archetypes. Especially for A1 and A2 no distinct maximum can be found. This is another artefact of the logarithmic scaling which we applied in the data pre- process procedure. Hence, by comparing the most enriched GO- groups of archetypes we can only identify several candidates for a biological process.

Instead of solely looking at the information contained in the coordinates of the identified archetypes, we therefore introduce two different measurements that help us to better identify the maximally enriched GO- groups. In order to find the right candidates among all GO- groups, we do not only look at the maximally enriched GO- expressions of the archetypes, but at the GO- expressions of all 451 samples. That is, we first compute the euclidean distance of all AML- samples to the archetypes and then, for each GO- expression, analyze the amount of expression, with respect to the distance from the archetype. Since the depiction of GO- expression of every single sample results in very jittery figures, we form bins of 10% of the data, 45 samples, and average the value of those. So for one GO-expression, the closest 45 samples get assigned one averaged value, the next 45 get assigned another value and so on. An example of this can be seen in fig. 8, upper row, where the same GO- expression, in this case *Reactome Endosomal Vacuolary Phathway*, has been selected for all archetypes and the amount of this expression is depicted with respect to each archetype respectively.

As maximally enriched at each respective archetype means, that the GO- expression must decay, as the distance to the archetype increases, our second measurement, to select GO- expression candidates from the high dimensional space, is a linear regression model. We evaluate the slope of each of the 242 GO- expressions with respect to each archetype. The GO- expressions with the most negative slope are then our candidates, most likely to correspond to a biological task of the AML- tumor. For each archetype, we depicted the GO- expression with the steepest negative slope in fig. 8, lower row. Note that for A2, this GO- expression is the same as in the upper row and that indeed *Reactome Endosomal Vacuolary Phathway*, which besides from A2 also happens to decay away from A0 is a feature that can be linked to the respective key task of A0 too.

More over it is interesting to cross compare the different enhanced features with respect to the three archetypes. Therefore a heat- map, fig. 9, has been created, where one can, for each archetype see, which of the GO- expressions are enhanced close to the archetype, i.e. have a very negative slope and which ones are neutral or increase with respect to the distance of the archetype.

In fig. 9 one can, for every archetype, A0,A1,A2, see which GO- expression shows a sharp decay with respect to the distance of the respective archetype and which not. Moreover does this figure reveal that the GO- expressions maximally enriched with respect to one archetype are minimally or neutrally enriched with respect to the other archetypes. This trend can specifically be observed when A1 and A2 are compared while A0 seems to be more of a generalist. Also fig. 8 upper row, gives rise to this hypothesis. *Reactome Endosomal Vacuolary Phathway*, the GO- expression that is maximally enriched at A2 is a gene-group, that also decreases, but weaker than others, with

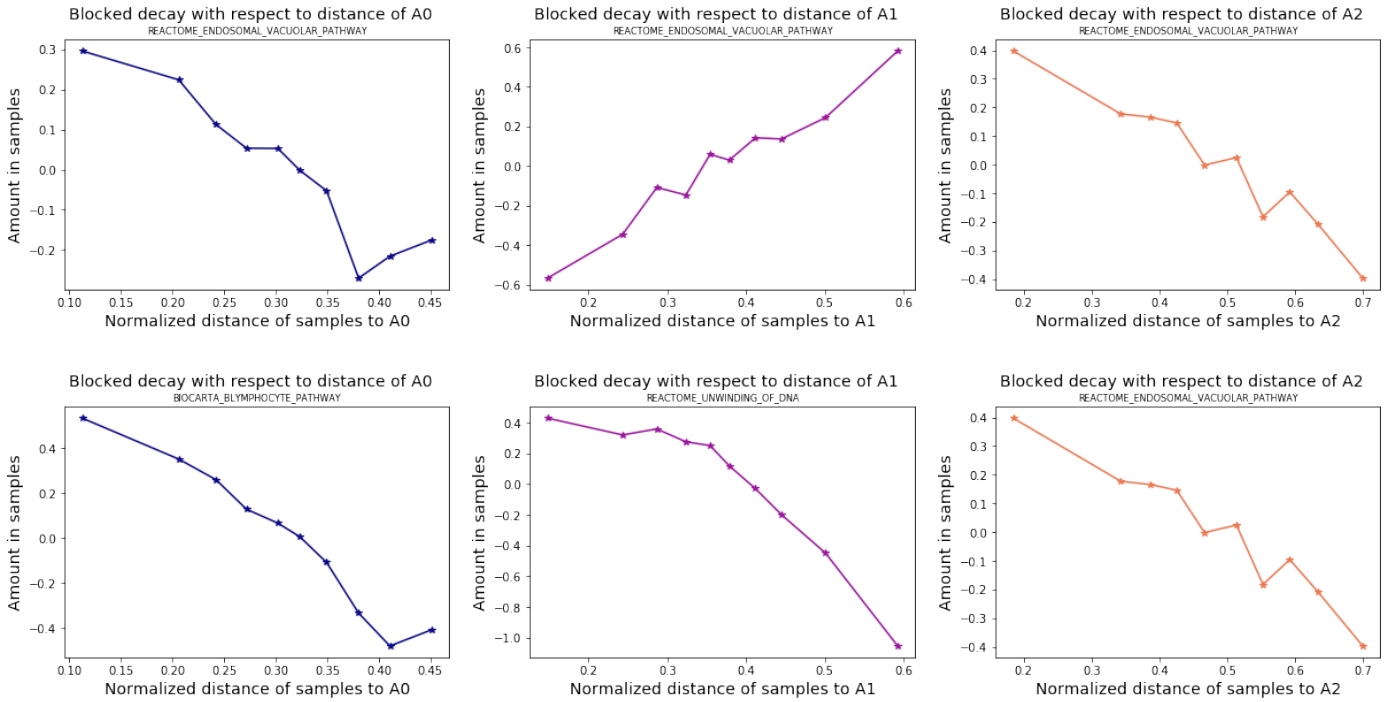

Figure 8: Key cancer features are maximally enriched at points nearest to the archetypes. We depict the average Go- expression amount for every 10% of samples with respect to their normalized euclidean distance to each archetype.
Upper row: Progression of the same GO- expression, *Reactome Endosomal Vacuolary Phathway* for all three archetypes.
Lower row: Selection of one GO- expression for each archetype, according to the maximal descent away from the archetype, i.e. most negative slope for linear regression. Suggesting that *Biocarta Blymphocyte Phathway*, *Reactome Unwinding of DNA* and *Reactome Endosomal Vacuolary Pathway* can be linked to key biological tasks.

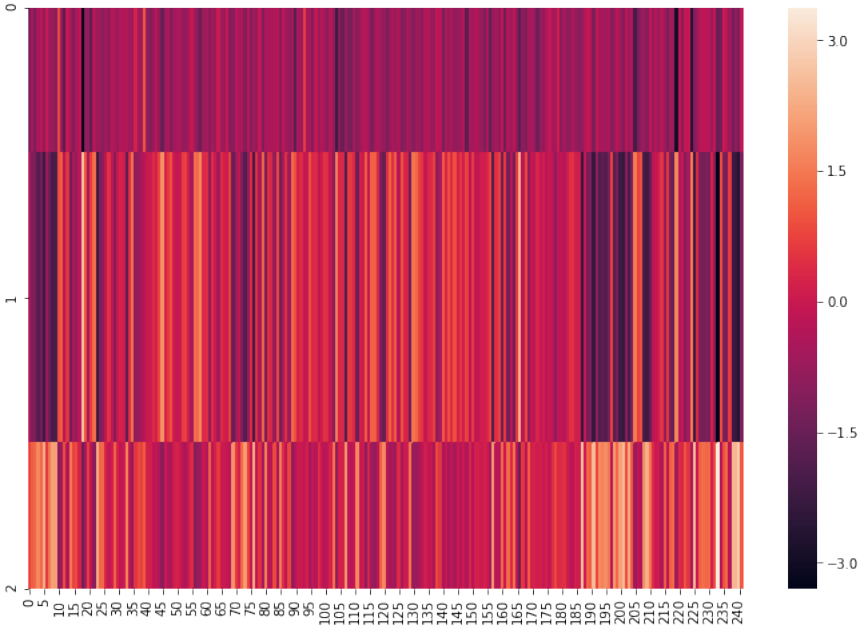

Figure 9: A heat- map revealing the decay strength of the full GO- expression spectrum away from the respective archetypes according to a linear model. Strongly decaying GO- expressions correspond to maximally enrichment at the archetypes and allow corresponding GO- expressions to be linked to key tumor cancer tasks. Dominant lines in A1 are less represented lines in A2, suggesting an evolutionary trade off scenario. A0 appears to be more of a generalist, however more distinct with respect to A1 and more related to A2.

respect to A0, however increases for A1.

The third measurement to identify the biological task of an archetype is the Wilcoxon Rank-Sum Test. This statistical test compares the likelihood that two data sets are originated from the same underlying distribution by ranking the data by its value. The sum of ranks of the two corresponding groups is compared and one can calculate a p-value: $p \in [0, 1]$. A large p-value indicates, that the null hypothesis, the data is drawn from the same underlying distribution, is true. A small p-value indicates the opposite. In our case we compare for every archetype A0, A1, A2 and every GO- group, how likely it is, that the average GO- expression amount in the first 10%, i.e first 45 samples closest to the respective archetype, is likely to be drawn from the same distribution than the rest of the data points. The GO- groups, for which its expression amount decays little with respect to the distance to the archetype show large p-values. Small p-values ($\approx 10^{-}21$) are found for Go-groups, for which the Go- expression amount either decays rapidly with respect to the distance to the archetype or increases with respect to the distance to the archetype. Both scenarios indicate that the GO- expression amount of the first 10 % of the samples are significantly different to the rest. The chance that the data is drawn from the same distribution is very unlikely in this case. Hence another qualitative indicator to identify a specific biological task is to find the GO- group with the smallest p-value for each archetype. Note that this indicator can only be used in combination with knowledge about the decay coefficient, previously found by fitting a linear regression model to the block averaged data, see fig. 8, because a small p-value does not indicate if the GO- expression amount grows or decays with respect to the distance to the archetype. Indeed, what we find is that the GO- groups with the strongest decay, with respect to the distance to the archetype, also happen to show small p-values.

Cross validating the three applied measurements, maximal GO- expression amount, slope with respect to distance and p-value, to identify a biological task for the three archetypes, we find that these independent measurements agree, see tab. 1. This means that the five most likely candidates of one measurement, happen to be within the five most likely candidates of the other measurements too. However, the ranking of the candidates under consideration might deviate from measurement to measurement.

Since by considering all samples and comparing the enrichment of the amount of GO- expression with respect to the distance of each archetype we find mutually exclusive expressions in the GO- spectrum, we can interpret the result within the framework of evolutionary trade- offs, as in [3] and [5][2].

[2]A more detailed table of sorted GO- expression candidates, according to the three measurements under consideration (expression in archetype itself, most negative slope and Rank-Sum Test) with their respective names can be found in the appendix.

5. Interpretation

The different genes expressed in tumors close to the different archetypes suggest 3 strategies employed by cancer cells to proliferate and survive.

A0 appears to divert the immune system's attention away from cancer cells. Close to A0, we find tumors which upregulate groups of genes needed to initiate immune response to infection micro-organisms such as IL8- signaling which recruits neutrophils and stimulates phagocytosis of invading micro- organisms, TOLL signaling which recognizes signals produced by micro- organisms, and genes involved in B- cell signaling.

Archetype 1 is a cell proliferation specialist. Close to that archetype, we find tumors which upregulate genes involved in replicating DNA, repairing DNA (mismatch repair, Fanconi pathway), in making proteins (tRNA synthesis, MTORC1 signaling), and in driving cell division (cell cycle).

Finally, archetype 2 may correspond to cancer cells resisting their own death by justifying their existence through a bogus mission of combating a non-existent infection. When cells become cancerous, built- in failsafe mechanisms normally causes them to suicide. But tumors close to archetype 2 have levels of CD40, TNF signaling and NFKb. NFkb is a molecule that causes cancer cells to resist death. Resisting death is normally needed in the context of fighting infection. NFKb is activated by signals such as TNF (Tumor Necrosis Factor) which bind different sensor proteins such as CD40. Binding of CD40 by TNF is thought to increase TNF production. This could lead to a vicious circle in which binding of CD40 by TNF keeps cancer cells alive through activation of NFKb and produces further TNF, leading more resistance to death, more TNF, and so on.

Of course, these strategies were inferred from the different gene programs are speculative and require further refining, both by deeper and careful review of the scientific literature on cancer and immunology, and by experimental follow-up. Nevertheless, the present analysis provides a starting point to explain the diversity of gene expression programs found here among acute myeloid leukemia (AML) tumors.

6. Conclusion

The main scope of this report is to infer biological task using machine learning techniques. After briefly analyzing AML- tumor samples with a K-means clustering method we move to a PCHA approach, which instead of finding typical representatives, aims to find distinct features in the data. We find this approach very well suited, since it can optimize with respect to mutually exclusive extrema and is therefore suited to explain evolutionary trade- offs biological specimen might face. In addition to deploying

the PCHA- algorithm do we further suggest three methods on how to select relevant candidates in the high dimensional Gene Onthology- space that correspond to biological tasks. Those tasks are then interpreted in order to explain the trade offs AML- tumor cells face.

Acknowledgement

We want to give special thanks to Jean Hausser [9] who did not only give us the opportunity to analyze the tumor data set, collected at the Weizmann Institute of Science, but also gave insightful feedback and suggestions with respect to the data analysis. Further could he, with his broad knowledge in both fields, span a coherent bridge between quantitative analysis and biological interpretation.

References

[1] Y. Jin, B. Sendhoff, Pareto-based multiobjective machine learning: An overview and case studies, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 38 (3) (2008) 397–415 (2008).

[2] A. Cutler, L. Breiman, Archetypal analysis, Technometrics 36 (4) (1994) 338–347 (1994).

[3] O. Shoval, H. Sheftel, G. Shinar, Y. Hart, O. Ramote, A. Mayo, E. Dekel, K. Kavanagh, U. Alon, Evolutionary trade-offs, pareto optimality, and the geometry of phenotype space, Science 336 (6085) (2012) 1157–1160 (2012).

[4] S. M. Keller, M. Samarin, M. Wieser, V. Roth, Deep archetypal analysis, CoRR abs/1901.10799 (2019). arXiv:1901.10799.
URL http://arxiv.org/abs/1901.10799

[5] Y. Hart, H. Sheftel, J. Hausser, P. Szekely, N. B. Ben-Moshe, Y. Korem, A. Tendler, A. E. Mayo, U. Alon, Inferring biological tasks using pareto analysis of high-dimensional data, Nature methods 12 (3) (2015) 233 (2015).

[6] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proceedings of the National Academy of Sciences 102 (43) (2005) 15545–15550 (2005).

[7] W. R. Fox, L. Kaufman, P. J. Rousseeuw, Finding groups in data: An introduction to cluster analysis., Applied Statistics 40 (3) (1991) 486 (1991). doi:10.2307/2347530.
URL https://doi.org/10.2307/2347530

[8] M. Mørup, L. K. Hansen, Archetypal analysis for machine learning and data mining, Neurocomputing 80 (2012) 54 – 63, special Issue on Machine Learning for Signal Processing 2010 (2012). doi:https://doi.org/10.1016/j.neucom.2011.06.033.
URL http://www.sciencedirect.com/science/article/pii/S0925231211006060

[9] J. Hausser, Jean Hausser Lab, Quantitative principles in tumor biology, https://www.hausserlab.org/, [Online; accessed 06-January-2020].

Appendix

In section 4 three different methods to select certain GO-groups that show maximally enrichment at the archetypes are described. That is we (1) look at the spectrum at the archetypes itself, (2) we consider the most significant decay of a feature away from the archetype and (3) we do a statistical evaluation with the Wilcoxon Rank-Sum Test. It is notable to say that the latter two methods suggest the same GO- expressions for each archetype. Furthermore can we see that qualitatively many of the highly expressed features at the archetypes, selection method (1), also appear within the top ranked features of method (2) and (3). However, since the latter two methods take all GO- groups of all sample points into consideration, they contain more information and are therefore prioritized by us.

Sorted by method (2) , in table 1 we present the 5 most important GO- group names, according to [6], with respect to each archetype. Inferring biological task from the tumor data set under consideration in section 5 has been done according to a complete form, ranking all 242, not only 5, GO- expressions.

GO- group	Decay-coeff.	P-Value	GO- expr.
Biocarta Blymphocyte Pathway	-3.29	1.3E-20	1.04
Reactome Signal Regulatory Protein SIRP Family Interactions	-2.95	4.73E-20	0.99
Reactome IKK Complex Recruitment Mediated by RIP1	-2.68	5.7E-18	0.88
SA-MMP Cytokine Connection	-2.26	1.63E-16	0.63
Reactome CD20 Dependent VAV1 Pathway	-2.09	7.35E-16	0.628
Reactome Unwinding of DNA	-3.34	1.52E-20	0.62
Reactome DNA Strand Elongation	-2.72	1.99E-20	0.55
Reactome POL Switching	-2.4	5.29E-20	0.53
Reactome Processive Synthesis on Lagging Strand	-2.39	5.29E-20	0.5
Reactome Extension of Teolmeres	-2.39	1.36E-19	0.48
Reactome Endosomal Vacuolar Pathway	-1.48	4.04E-24	0.61
Biocarta Blymphocite Pathway	-1.39	6.21E-24	0.5
Biocarta NTHI Pathway	-1.1	9.42E-24	0.47
KeGG Circadian Rhythm Mammal	-0.99	1.22E-23	0.39
Biocarta CD40 Pathway	-0.98	1.75E-23	0.39

Table 1: Ranked importance, descending order, of GO- group for each archetype, A0,A1,A3. The ranking is performed according to the strongest decay of the respective feature away from the archetype.