Progression analysis of disease applied to breast cancer research

Rachel Jeitziner
under the supervision of Kathryn Hess Bellwald and Cathrin Brisken
École Polytechnique Fédérale de Lausanne
rachel.jeitziner@epfl.ch

Abstract In this paper, we present several applications of a recently developed mathematical field called topological data analysis (TDA). The poster focusses on two different methods of TDA : Progression Analysis of Disease and the analysis of Betti numbers. We apply these techniques to a set of microarrays from tissue donated by women undergoing mammoplasty surgery. These are results from breast cancer research ; obtained under varying experimental conditions. We find that topological features describe significant structures of the data, insights that could not be gained with standard tools.

First application : RANKL protein

Cancer arises from cells that leave the cell cycle and start to proliferate in an uncontrolled manner. This proliferation could be induced by hormones that are impinging on the breast. Researchers found that receptor activator of nuclear factor κB ligand (RANKL) is a protein involved in progesterone induced proliferation. To understand what this protein is inducing, **we investigate genes that are differentially expressed, when one stimulates healthy human tissue with RANKL.**
From the same patient, we stimulate one tissue with RANKL and another one with vehicle (that is a solution in which the RANKL protein is put into suspension).
The standard statistical test (a moderated t-test), highlights only 11 genes whose expression is significantly altered between the two groups.
Using instead progression analysis of disease (PAD)[2], with the unstimulated samples as the healthy state model (HSM) (the "normal group"), we are able to measure the deviation from the HSM for every stimulated sample, implying that we can directly give a **qualitative measure** of how much a sample is "changed" upon stimulation with RANKL, where blue means close to the HSM and red means far away.

PAD helped us reveal that :
- Sample 6 is the closest to normal. This woman was the youngest, and had the lowest level of serum progesterone.
- Sample number 2 and 3, which had both a high level of serum progesterone, are altered across many genes from the HSM.
- Sample 3 had a lower level than sample 2, however she was taking oral contraceptives and was the oldest patient.

$\Rightarrow$ Age and the contraceptive methods have an influence on the activity of RANKL, but also the level of serum progesterone.

⋆ **TDA works better than usual statistical methods!**
⋆ **Not all samples react the same way upon stimulation with RANKL!**
⋆ **Almost 5'000 significant genes (whereas 11 with the standard method)!**

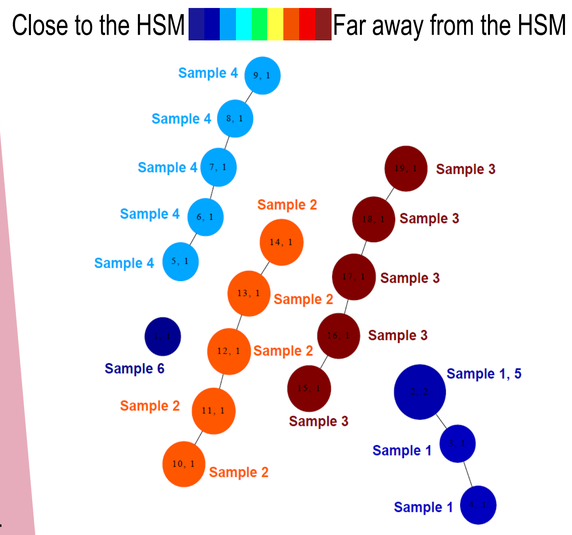


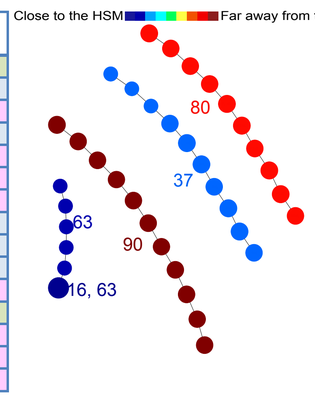
Fig. 1: PAD output, 9 intervals with 80% of overlap, the chosen filter function is $f(V) = (\Sigma_r \mid g_r \mid^5)^{4/5}$, where V is one of the diseased component ($Dc.T$) obtained by DSGA and g_r are the entries of the vector $Dc.T$. In this case r ranges from 1 to 4958.

Second application : Menstrual cycle

Since hormones are implicated in tumour formation and/or growth, **we are interested in the variations in gene expression throughout the menstrual cycle**. Knowing that progesterone may be a major driver of cancer, we are mostly interested in seeing the differences between samples from women in the luteal phase (characterized by a progesterone peak) compared to the follicular phase (characterized by a period of quiescence of progesterone). Moreover, it is known that women taking oral contraceptives (OC) are at a higher breast cancer risk. Explaining these facts is of major interest, since contraceptives are suppose to mimic the luteal phase. Therefore, comparing the tissue of women taking OC to tissue from women in the luteal phase would show how different their gene profiles are. Hence, microarrays were performed on several different tissues, from women either in the follicular phase, or in the luteal phase, or taking oral contraceptives.

What we found with PAD is that women 80 and 90 taking OC have a very different gene profile than women in the luteal phase (which here is the HSM!).

Number	Progesterone Level ng/mL	Age	Follicular or Luteal	Hormonal Contraception
5	1.1	45	F	-
11	37.2	33	L	-
16	0.6	28	-	Meliane, Meloden 21, Diane 35, 12 years
22	18.6	39	L	-
27	14.8	32	-	Yes, during 10 years
29	0.5	36	-	Minulet, during 10 years
37	1.1	34	-	Yasmine, 6 months
39	26.3	44	L	-
46	17.4	29	L	-
51	19.4	35	L	-
63	1.6	41	-	Yes
69	0.4	35	F	-
80	0.4	31	-	Marvelon
83	23.2	43	-	Yes, 10 years
90	0	41	-	Desarex, 15 years



- Strange and interesting fact : both (80, 90) are taking contraceptives containing desogestrel.
- This progestin has been associated with a high risk of thromboembolism!

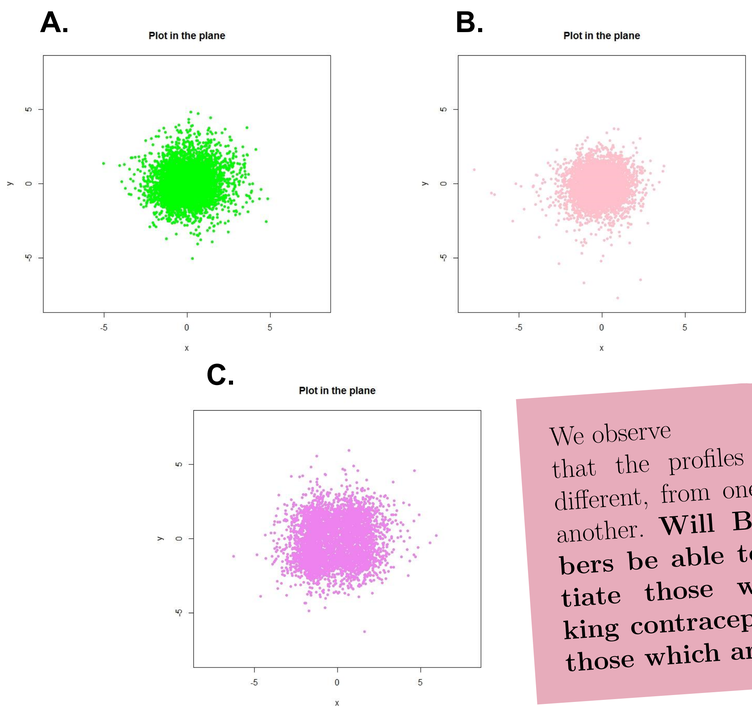
Fig. 2: Patient information and PAD output of myoepithelial cells, the healthy state model is defined by women in the luteal phase, the diseased vectors are the women taking oral contraceptives, 9 intervals with 90% of overlap, the chosen filter function is $f(V) = (\Sigma_r \mid g_r \mid^5)^{4/5}$.

Topology distinguished between contraceptives that behave like the luteal phase and those which generate an extremely different profile! Again, no significant genes were found with standard tools, unlike PAD which revealed thousands of genes.

Third application : Menstrual cycle revisited

In our previous quest to find differentially expressed genes during the phases of the menstrual cycle, we saw differences from one person to another. After reading the paper of J. Arsuaga et al.[1], we decided to apply these ideas to the setting of microarrays. We could not directly apply their methods, since we had microarray data and they had comparative genomic hybridization array data. In other words :
- values approaching 0 for genes that are not changed
- high values for genes that changed compared to a "reference population".

However, we observed that we also have the same setting after application of disease-specific genomic analysis (DSGA), which is the first statistical part of the progression analysis of disease method $\Rightarrow$ **we could try to apply the same method, on the DSGA results.** Hence, we want to see differences between women taking oral contraceptives and women in the luteal phase, after comparing them with DSGA to the follicular phase. Again, since oral contraceptives are supposed to mimic the luteal phase, we expect that their gene profile is close that of women in the luteal phase.



We observe that the profiles seem very different, from one woman to another. **Will Betti numbers be able to differentiate those women taking contraceptives from those which are not ?**

Fig. 3: Plot of the vectors obtained after DSGA method, with women in the follicular phase representing the HSM. A. Sample 39, beginning of luteal phase. B. Sample 46, end of luteal phase. C. Sample 80, taking oral contraceptives.

- Calculating the 0-th Betti number revealed differences regarding contraceptives.
- Again (!!) sample 80 had a more complex gene profile than the others.
- There are some contraceptives with profiles similar to those of women in the luteal phase.

This provokes the question : Is there a possible **classification** of contraceptives with the help of Betti numbers, which could perhaps help explaining which contraceptives are harmful to women's breasts ?

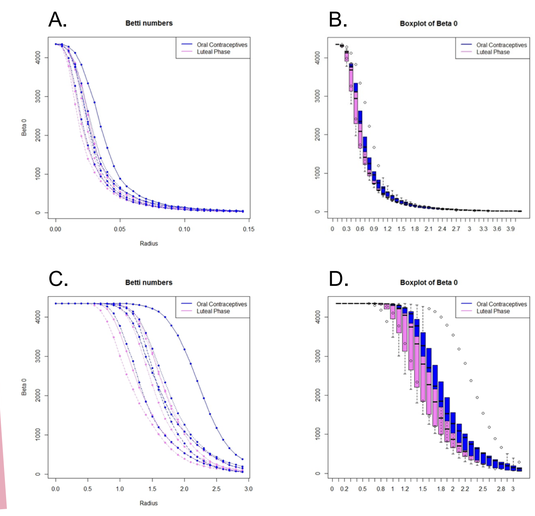


Fig. 4: A. β_0 computed for dimension n = 3. B. Boxplot of β_0 values at each step for those two groups, n=3. C. β_0 computed for dimension n = 10. D. Boxplot of β_0 values at each step for those two groups.

Conclusion

Progression Analysis of Disease highlights genes that are significantly differentially expressed even if it is just for a small number of patients. In particular, we observe that the activity of RANKL is probably linked to the age of the patient, the contraceptive history, and importantly, the level of serum progesterone. On the other hand, Betti numbers reveal notions of complexity of a gene profile in comparison with a reference group, using DSGA. All these results could not be found with standard tools. This method was inspired by the work of J. Arsuaga et al.

Method

All PAD computation were done using the software on the webpage www.comptop.stanford.edu/pad/. All other computations were either realised using the R project version 3.0.2 or the software Perseus for computing simplicial complexes.

References

1. J. Arsuaga, *et al.*, 2010, *Applications of computational homology to the analysis of treatment response in breast cancer patients*, Topology and its Applications, 157, 157–164.
2. M. Nicolau, *et al.*, 2011, *Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival*, Proc Natl Acad Sci USA, 108(17) 7256-70.