Skin Disease Basic Analysis

Analysing gene expresseion for Psoriasis and Atopic Dermatitis

T.P.A. Beishuizen (0791613)
Biomedical Engineering - Computational Biology
Computer Science - Data Mining
Eindhoven, University of Technology
Email: t.p.a.beishuizen@student.tue.nl

March 5, 2018

Contents

1 Introduction

At the Computational Biology department (cBio) of Biomedical Engineering (BME), many requests are made to analyse gathered data. This data usually stems from research in hospitals, but can also be from other BME groups and publicly available data. Currently a standard is missing to efficiently analyse those data sets. With the vast number of data sets that are available, such a standard in the form of a framework on data analysis would be valuable. It would speed up projects and give them a higher chance to succeed the goal, due to improved efficiency.

An example of biomedical data sets was based around gene expression of skin diseases. Two skin diseases were tested, psoriasis and atopic dermatitis, the latter one better known as a form of eczema. The expression of a big number of genes was tested for skin disease patients on skin affected by the disease (lesional skin) and skin not affected by the disease (non-lesional skin). At last there were normal patients, that did not suffer from the skin disease. Nine data sets were available for these data set, six for psoriasis and three for atopic dermatitis. The number of tested skin plaques ranged from 28 to 180 whereas the number of tested genes is the same for every set, namely 54675.

2 Skin Diseases Data Sets

Skin diseases could form a major disability in someone's life. Whereas skin diseases were not as life threatening as diseases such as Cancer, Alzheimer and AIDS, the could lower quality of life significantly. When looking at health-related quality of life (HRQL), patients with psoriasis showed same problems as patients with other major chronic health conditions.[1] Patients with both psoriasis and eczema suffered from severe itching symptoms and possibly even severe pains. Further insights in these skin diseases could help alleviate their unwanted side-effects and help improve the patients' quality of life.[2]

Information on both of these skin diseases could be found from nine data sets from the NCBI database[11]. The data sets consisted of information on skin of patients with the disease (lesional skin) and skin without the disease (non-lesional skin). In several experiments this skin was taken from the same patient. Also some skin was taken from patients not suffering from the diseases at all. Six data sets focused on Psoriasis and three focused on atopic dermatitis. These data sets consisted of a total number of 54675 features, each of them corresponding to a specific gene. Not many skin samples were taken, ranging from 28 to 180. Also, since every data set was created by different people, some minor differences were present in them as well (Table 1).

The nine data sets are rich in information. The dimensionality is very high and if combined also houses a decent number of samples. Several challenges arise in the data set, too, as biomedical data sets often have. Three of these challenges are discussed for this case.

At first the challenge of handling nine different data sets was important. Even though the sets were created based on the NCBI database[11], the layouts were not identical. These difference originated from the intended research goals and the data availability. It is not possible to just concatenate samples without some form of preprocessing. Only the parts that are the same all over the data sets should be taken and all other parts omitted. A first look would be best on the nine data sets separately so initial ideas found with as less bias from combining them as possible.

A second challenge could be found in the high number of features. There were 74675 featured measured, averaged a 1000 times the number of samples. The genes that actually were significantly involved in the skin diseases however should be about $1/1000^{th}$ of the total number of measured genes. Many features should be redundant and removed during preprocessing, a valuable and complex step in biomedical data mining.

The third challenge was about data volume. The number of samples differed from 28 to 180, all of them being a very low number compared with the number of features. This indicates that the number of samples represent the complete sample space poorly, not clearly showing the boundaries between the areas. This could create problems during machine learning with such a low training and test set. Several cases will arise where accidentally all training and test set agree with the

Table 1: Details of the nine skin disease data sets. The number of samples and features has been given, as well as remarks of the skin types.

Disease	Data set name	Sample size	Features	Remarks
Psoriasis	**GSE13355** [3]	180	54676	Three skin types: - NN (normal, 64 samples) - PN (non-lesional, 58 samples) - PP (lesional, 58 samples)
	GSE30999 [4]	170	54676	- No normal patients - Non-lesional (85 samples) - Lesional (85 samples)
	GSE34248 [5]	28	54676	- No normal patients - Non-lesional (14 samples) - Lesional (14 samples)
	GSE41662 [5]	48	54676	- No normal patients - Non-lesional (24 samples) - Lesional (24 samples)
	GSE78097 [6]	33	54676	Different types of skin samples: - Normal (6 samples) - Mild Psoriasis (14 samples) - Severe Psoriasis (13 samples)
	GSE14905 [7]	82	54676	- Normal skin (21 samples), - Non-lesional skin (28 samples) - Lesional skin (33 samples)
Atopic Dermatitis	**GSE32924** [8]	33	54676	- Normal skin (8 samples) - Non-lesional skin (12 samples) - Lesional skin (13 samples)
	GSE27887 [9]	35	54676	Different type of skin samples, pre and post treatment of skin: - Pre non-lesional (8 samples) - Post non-lesional (9 samples) - Pre lesional (9 samples) - Post lesional (9 samples)
	GSE36842 [10]	39	54676	Also difference between acute and chronic dermatitis. - Normal (15 samples) - Non-lesional (8 samples) - Acute lesional (8 samples) - Chronic lesional (8 samples)

algorithm, whereas other samples from the sample space would not.

2.1 Additional Data

The genes were the same for all of these nine different data sets. The NCBI database[11] also provided separate data containing various information for every gene. This information included gene ID, commonly known name and abbreviation and which gene bank it originated from. Aside from this general knowledge it also contained processes the gene was involved with, cellular locations of the gene as well as molecular reactions they are involved in. This data could be used to find relations between several processes and their corresponding genes.

3 Methods

Several techniques were used to achieve

3.1 Feature Reduction

3.2 Clustering

3.3 Psoriasis Versus Atopic Dermatitis

4 Results

Results introduction

4.1 Feature Reduction

4.2 Clustering

4.3 Psoriasis Versus Atopic Dermatitis

5 Conclusion

6 Discussion for Framework

References

[1] S. R. Rapp, S. R. Feldman, M. L. Exum, A. B. Fleischer, and D. M. Reboussin, "Psoriasis causes as much disability as other major medical diseases," *Journal of the American Academy of Dermatology*, vol. 41, no. 3, pp. 401–407, 1999.

[2] S. Jowett and T. Ryan, "Skin disease and handicap: an analysis of the impact of skin conditions," *Social science & medicine*, vol. 20, no. 4, pp. 425–429, 1985.

[3] R. P. Nair, K. C. Duffin, C. Helms, J. Ding, P. E. Stuart, D. Goldgar, J. E. Gudjonsson, Y. Li, T. Tejasvi, B.-J. Feng, *et al.*, "Genome-wide scan reveals association of psoriasis with il-23 and nf-κb pathways," *Nature genetics*, vol. 41, no. 2, pp. 199–204, 2009.

[4] M. Suárez-Farinas, K. Li, J. Fuentes-Duculan, K. Hayden, C. Brodmerkel, and J. G. Krueger, "Expanding the psoriasis disease profile: interrogation of the skin and serum of patients with moderate-to-severe psoriasis," *Journal of Investigative Dermatology*, vol. 132, no. 11, pp. 2552–2564, 2012.

[5] J. Bigler, H. A. Rand, K. Kerkof, M. Timour, and C. B. Russell, "Cross-study homogeneity of psoriasis gene expression in skin across a large expression range," *PLoS One*, vol. 8, no. 1, p. e52242, 2013.

[6] J. Kim, R. Bissonnette, J. Lee, J. C. da Rosa, M. Suárez-Fariñas, M. A. Lowes, and J. G. Krueger, "The spectrum of mild to severe psoriasis vulgaris is defined by a common activation of il-17 pathway genes, but with key differences in immune regulatory genes," *Journal of Investigative Dermatology*, vol. 136, no. 11, pp. 2173–2182, 2016.

[7] Y. Yao, L. Richman, C. Morehouse, M. De Los Reyes, B. W. Higgs, A. Boutrin, B. White, A. Coyle, J. Krueger, P. A. Kiener, *et al.*, "Type i interferon: potential therapeutic target for psoriasis?," *PloS one*, vol. 3, no. 7, p. e2737, 2008.

[8] M. Suárez-Fariñas, S. J. Tintle, A. Shemer, A. Chiricozzi, K. Nograles, I. Cardinale, S. Duan, A. M. Bowcock, J. G. Krueger, and E. Guttman-Yassky, "Nonlesional atopic dermatitis skin is characterized by broad terminal differentiation defects and variable immune abnormalities," *Journal of Allergy and Clinical Immunology*, vol. 127, no. 4, pp. 954–964, 2011.

[9] S. Tintle, A. Shemer, M. Suárez-Fariñas, H. Fujita, P. Gilleaudeau, M. Sullivan-Whalen, L. Johnson-Huang, A. Chiricozzi, I. Cardinale, S. Duan, *et al.*, "Reversal of atopic dermatitis with narrow-band uvb phototherapy and biomarkers for therapeutic response," *Journal of Allergy and Clinical Immunology*, vol. 128, no. 3, pp. 583–593, 2011.

[10] J. K. Gittler, A. Shemer, M. Suárez-Fariñas, J. Fuentes-Duculan, K. J. Gulewicz, C. Q. Wang, H. Mitsui, I. Cardinale, C. de Guzman Strong, J. G. Krueger, *et al.*, "Progressive activation of t h 2/t h 22 cytokines and selective epidermal proteins characterizes acute and chronic atopic dermatitis," *Journal of Allergy and Clinical Immunology*, vol. 130, no. 6, pp. 1344–1354, 2012.

[11] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: Ncbi gene expression and hybridization array data repository," *Nucleic acids research*, vol. 30, no. 1, pp. 207–210, 2002.