Large-Scale Sentence Clustering from Electronic Health Records for Genetic Associations in Cancer

M. F. Pradier, F. Perez-Cruz[*]
Department of Signal Theory
and Communications
Univ. Carlos III in Madrid, Spain
melanie@tsc.uc3m.es

S. Stark, S. Hyland, J.E. Vogt, G. Rätsch
Department of Computational Biology
Memorial Sloan-Kettering
Cancer Center, New York, USA
starks@cbio.mskcc.org

Abstract

This paper proposes a new framework to find associations between somatic mutations and clinical features in cancer. The clinical features are directly extracted from the Electronic Health Records by performing a large-scale clustering of the sentences. Using a linear mixed model, we find significant associations between EHR-based phenotypes and gene mutations, while correcting for the cancer type as a confounding effect. To the author's knowledge, this is the first attempt to perform genetic association studies using EHR-based phenotypes. Such research has the potential to help in the discovery of unknown mechanisms in cancer, which will allow to prevent the disease, monitor patients at risk, and design tailored treatments for the patients.

1 Introduction

Cancer encompasses not one, but a vast group of genetic diseases that are not very well understood yet. In the last decade, high-throughput genotyping technologies have led to the discovery of cancer-correlated gene mutations, most of which were not previously suspected to be related to carcinogenesis [1]. However, only the gene mutations with very strong effects have been discovered and many other genes with weaker effects still remain to be found [2]. Genetic-association studies have been widely used in the search for such genes, but success has been limited so far [3].

A first difficulty in cancer association studies is the immense phenotypic heterogeneity. Complex diseases such as melanoma, prostate, or breast cancer have indeed proved to be very heterogeneous, which reduces statistical power in the discovery method and causes some associations to remain hidden [4, 5, 6]. As a result of such diversity, most cancer phenotypes are still poorly understood, and are extremely hard to define accurately, even using ontologies. Also, given the huge number of phenotypical variants, conducting valid association studies in rare cancers is particularly challenging due to inadequate cohort sizes [2]. In such a scenario, new approaches for genetic association studies must be found that are able to capture phenotype heterogeneity.

In recent years, the adoption of Electronic Health Records (EHR) in hospitals has increased dramatically, and has become an interesting resource for phenotyping [7, 8]. EHR consist of both structured and unstructured information. Although structured data is very valuable, most of the clinical data, e.g., around 98% of the EHRs, comes as unstructured notes [9]. These include a broad spectrum of clinically-relevant phenotypic information, and might be useful to identify new phenotypes, still unclassified in ontologies [8, 9].

This paper presents a new framework to find associations between somatic mutations and clinical features in cancer, that deals with phenotype heterogeneity, rare cancers, and the lack of precise

[*] Also Machine Learning Scientist at Bell Labs.

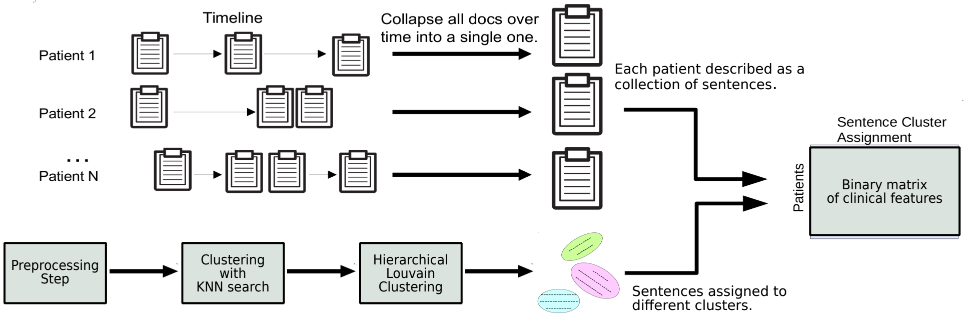

Timeline
Patient 1
Patient 2
...
Patient N
Collapse all docs over time into a single one.
Each patient described as a collection of sentences.
Sentence Cluster Assignment
Patients
Binary matrix of clinical features
Preprocessing Step
Clustering with KNN search
Hierarchical Louvain Clustering
Sentences assigned to different clusters.

Figure 1: **Process to obtain the phenotype information.** The upper part shows how we generate a summary document for each patient. The sentence clustering procedure is shown at the bottom.

phenotypical terms in ontologies. The considered clinical features are directly extracted from the unstructured EHRs of the patients at the sentence level. We look at sentences because these are small entities, but still precise enough to describe homogeneous sub-phenotypes. Since each sentence is most probably unique, we need to perform a large-scale clustering previous to the analysis. The resulting features are then used in a genetic association study against somatic mutations, using a linear mixed model. We consider multiple cancers jointly in order to increase statistical power, allow for the analysis of rare cancers, and identify shared mechanisms across cancers. Indeed, most cancers are known to share a common pathogenesis despite specificities of the cell type and tissue origin [10]. In our study, we use the cancer type as a confounding factor. This avoids getting already well-known per-cancer mutations, and are able to obtain less well-known associations with gene mutations of smaller effect size. Our procedure generates potentially interesting associations that might be useful for further research in oncology.

2 Feature Extraction

2.1 Large-scale Sentence Clustering from EHR

This Section describes our procedure to obtain a $P \times Q$ matrix Y of phenotypes, where Q designs the extracted clinical features, and P is the number of patients. We use our method to partition sentences from 2,008,462 EHRs provided by 295,154 cancer patients of the Memorial Sloan-Kettering Cancer Center in New York. Figure 1 describes the main elements of our approach. Each patient has a list of EHRs associated, which we collapse into a single big document (upper part). We thus obtain a patient by sentences matrix, most of which are unique sentences. In order to be able to use the sentences as clinical features, we perform a large-scale clustering (bottom part) in three steps.

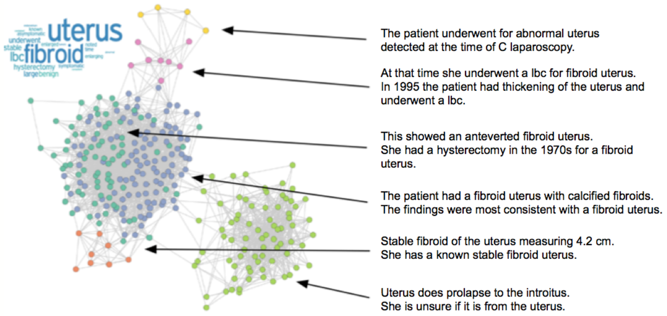
First, we perform a basic parsing of the clinical notes. In particular, we remove all stop words and the most frequent words, as these do not contribute to the overall meaning of the sentences. We also ignore punctuation, replace upper case letters by lower case ones, and collapse all numericals into a single pound sign (#). From the original 101,564,226 sentences, 34,612,424 unique sentences remain. These sentences are stored as a sparse matrix of sentences by words.

The second step consists in finding the K-nearest neighbors (KNN) for each sentence S_i. For that, we need to compute the similarity of each pair of sentences, which will depend on the number of shared words. Intuitively speaking, words should not contribute equally to the similarity measure. For example, a rare specific medical term should have a bigger contribution than a more common word. To capture this effect, words are weighted by the log of their inverse frequency. If S is the set of all unique sentences, the weight f_w of a term w is given by:

$$f_w \equiv f(w) = -\log \frac{|\{S_i : S_i \in S, w \in S_i\}|}{|S|} \tag{1}$$

We define a weighted jaccard index ϕ as a similarity measure:

$$\phi(S_i, S_j) = \frac{\sum_{w \in S_i \cap S_j} f(w)}{\sum_{w \in S_i \cup S_j} f(w)} \tag{2}$$



The patient underwent for abnormal uterus detected at the time of C laparoscopy.

At that time she underwent a lbc for fibroid uterus. In 1995 the patient had thickening of the uterus and underwent a lbc.

This showed an anteverted fibroid uterus. She had a hysterectomy in the 1970s for a fibroid uterus.

The patient had a fibroid uterus with calcified fibroids. The findings were most consistent with a fibroid uterus.

Stable fibroid of the uterus measuring 4.2 cm. She has a known stable fibroid uterus.

Uterus does prolapse to the introitus. She is unsure if it is from the uterus.

Figure 2: **Partial Visualization of the Louvain's method output.** Each node is a single sentence. Colors of the nodes correspond to the Louvain's most granular clustering. The entire picture is the second most granular clustering.The word cloud is created with the vocabulary of the high-level cluster. Size is determined from frequency of the term. A few sample sentences are shown.

Since the sentence by word matrix is extremely sparse, the computational load decreases considerably by only computing the similarity of sentences sharing at least one word [1].

In the third step, the top K neighbors are stored in a dense matrix. For each sentences neighbor, we compute the number of neighbors they share, take the top K neighbors and store these as an even sparser matrix. Such procedure is often done in social networks to remove noise and detect communities [11]. Finally, this matrix is treated as an adjacency matrix of a network, where nodes are sentences and edges are weighted by the similarities. Sentences are then clustered using the Louvain method [11]. This is an unsupervised hierarchical clustering method designed to detect community structure in networks. It works by iteratively and greedily optimizing local modularity. The algorithm produces multiple levels of clustering, with lower-level clusters acting as more detailed, granular clusters than larger higher-level clusters. Figure 2 shows one of the resulting clusters from the method.

2.2 Genetic Information from the MSK IMPACT database

So far, genomic testing of tumors has been done routinely only for certain solid cancer tumors, such as melanoma, lung, or colon cancer. For most cancers, the available tests have been limited to analyzing one or a handful of genes at a time, and within each gene, only the most common mutations could be detected [12]. A new targeted tumor sequencing test called MSK-IMPACT (Integrated Mutation Profiling of Actionable Cancer Targets) is able to detect gene mutations and other critical genetic aberrations in both rare and common cancers [13]. This test is a hybridization capture-based next-generation sequencing assay for targeted deep sequencing of all exons and selected introns.

For each patient, tumor cells are compared with healthy cells of that same patient, extracted from the blood stream. In this paper, a gene is said to be mutated when there exists at least one difference in the sequence between the tumor cells and healthy cells for that particular gene[2]. The resulting information is summarized in a binary matrix X with dimensions $P \times G$, where P is the number of patients, and G is the number of genes considered by the test, i.e., 342 genes. These genes have been shown to play a role in the development or behavior of tumors, although their individual relation to specific phenotypes remains obscur. [13, 14].

3 Discovery Method: Linear Mixed Model

Linear mixed models are one of the most widely used approaches in genetic association studies due to its capacity to account for confounding effects and limit the number of false associations [15, 16]. Let x_{pg} be the indicator variable for a somatic mutation in gene $g \in \{1, ...G\}$ and patient

[1] For that reason, removing commonly occurring words vastly improves the run time.

[2] The considered sequencing technology is able to remove most of the technical noise, which is not often the case with other technologies.

$p \in \{1, ...P\}$. This is a binary variable that will be one if any somatic mutation occurred in the corresponding gene. Let y_{pq} be the binary indicator variable of the clinical feature $q \in \{1, ...Q\}$ for a given patient p. Finally, let us define c_{pl} as the binary assignment variable of patient p to the cancer type $l \in \{1, ...L\}$. For each pair of gene g and clinical feature q, we train a linear mixed model, as follows:

$$y_{.q} = x_{.g}\beta + u_q + \varepsilon_q \tag{3}$$

where $\beta \in \mathbb{R}^{G\times 1}$ are the fixed effects, $u \in \mathbb{R}^{P\times 1}$ are the random effects, and $\varepsilon \in \mathbb{R}^{P\times 1}$ is the observational noise. The prior assumptions for the structured and uniform noises are $u \sim \mathcal{N}(0, \sigma_u^2 K)$ and $\varepsilon \sim \mathcal{N}(0, \sigma_e^2 I)$, K refers to a similarity matrix between the patients computed as the cosine distance of the vectors $c_{p_i.}$ and $c_{p_j.}$ of confounders, i.e., the cancer type. In other words, $K = CC^T$. The linear mixed-model assumes that the output $y_{.q}$ is Gaussian-distributed. Since our data is binary and count data, we apply a standard rank-based inverse normal transformation beforehand [17]. Thus, we have

$$y_{.q} \sim \mathcal{N}(x_{.g}\beta, \sigma_u^2 K + \sigma_e^2 I) \tag{4}$$

The model parameters are found by maximizing the log likelihood using standard optimization techniques within a python platform called LIMIX [16]. In the final step, we obtain p- and q-values for each pair $(y_{.q}, x_{.g})$ using likelihood ratio tests. If H_0 is the null hypothesis, t_i the statistic of test i and α the significance level, p-value= $\Pr(t_i \geq \alpha | H_0)$ and q-value= $\Pr(H_0 | t_i \geq \alpha)$. Under the Gaussian assumption, we know the theoretical null hypothesis distribution, namely, that the likelihood ratio statistics follow a χ^2 distribution. Finally, we perform a correction for multiple hypothesis testing using Bonferoni correction [18].

4 Results and conclusion

Our association study is performed using 2205 patients (those having both genetic and clinical information), 343 genes, 7k clinical features, and 63 different cancer types. After correcting for confounders, we find 332 significant hits having a q-value below 1%. Figure 3 shows some association examples found by our model. Some are already well known, like gene APC with colon cancer. Other associations such as TRAF7 with cystic carcinoma have been confirmed very recently, and others like ALK with anemia are still just hypothetical in the medical community. Interestingly, we are also able to find associations with very low effect size such as ERBB4 with colon carcinoma, which was not confirmed until this year. A more extensive list of associations can be found in the Supplementary. In general, our method is able to find potentially interesting associations, with many of them having been confirmed within the last decade. We hope that this procedure will help physicians and biologists to better understand the underlying mechanisms of cancer.

References

[1] D. F. Easton and R. A. Eeles, Human Molecular Genetics **17**, R109 (October 2008).

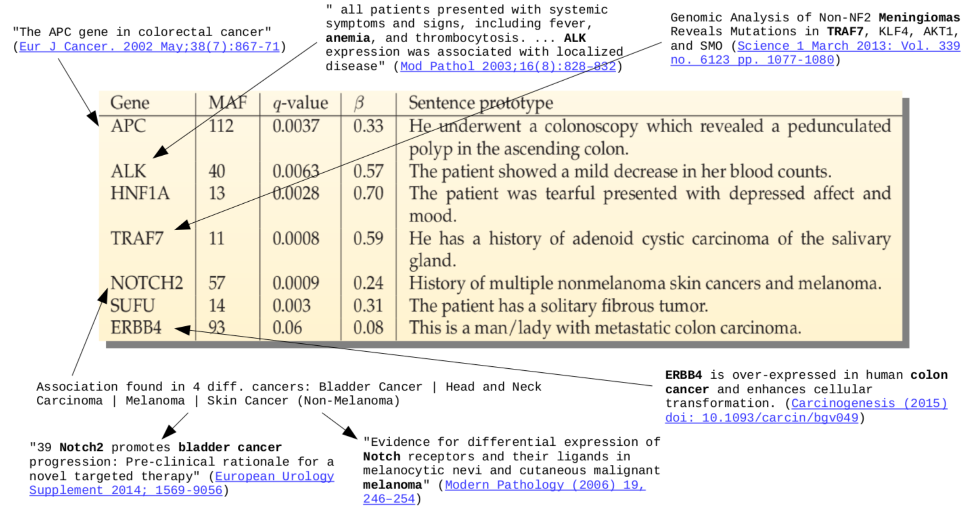
[2] U. Andersson, R. McKean-Cowdin, U. Hjalmars and B. Malmer, Acta Oncologica (Stockholm, Sweden) **48**, 948 (2009).

[3] P. D. P. Pharoah, A. M. Dunning, B. A. J. Ponder and D. F. Easton, Nature Reviews Cancer **4**, 850 (November 2004).

[4] E. Quintana, M. Shackleton, H. R. Foster, D. R. Fullen, M. S. Sabel, T. M. Johnson and S. J. Morrison, Cancer Cell **18**, 510 (November 2010).

[5] E. Lalonde, A. S. Ishkanian, J. Sykes, M. Fraser, H. Ross-Adams, N. Erho, M. J. Dunning, S. Halim, A. D. Lamb, N. C. Moon, G. Zafarana, A. Y. Warren, X. Meng, J. Thoms, M. R. Grzadkowski, A. Berlin, C. L. Have, V. R. Ramnarine, C. Q. Yao, C. A. Malloff, L. L. Lam, H. Xie, N. J. Harding, D. Y. F. Mak, K. C. Chu, L. C. Chong, D. H. Sendorek, C. P'ng, C. C. Collins, J. A. Squire, I. Jurisica, C. Cooper, R. Eeles, M. Pintilie, A. Dal Pra, E. Davicioni, W. L. Lam, M. Milosevic, D. E. Neal, T. van der Kwast, P. C. Boutros and R. G. Bristow, The Lancet Oncology **15**, 1521 (December 2014).

[6] M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl and J. H. Moore, American Journal of Human Genetics **69**, 138 (July 2001).



"The APC gene in colorectal cancer" (Eur J Cancer. 2002 May;38(7):867-71)

" all patients presented with systemic symptoms and signs, including fever, **anemia**, and thrombocytosis. ... **ALK** expression was associated with localized disease" (Mod Pathol 2003;16(8):828–832)

Genomic Analysis of Non-NF2 **Meningiomas** Reveals Mutations in **TRAF7**, KLF4, AKT1, and SMO (Science 1 March 2013: Vol. 339 no. 6123 pp. 1077-1080)

Gene	MAF	q-value	β	Sentence prototype
APC	112	0.0037	0.33	He underwent a colonoscopy which revealed a pedunculated polyp in the ascending colon.
ALK	40	0.0063	0.57	The patient showed a mild decrease in her blood counts.
HNF1A	13	0.0028	0.70	The patient was tearful presented with depressed affect and mood.
TRAF7	11	0.0008	0.59	He has a history of adenoid cystic carcinoma of the salivary gland.
NOTCH2	57	0.0009	0.24	History of multiple nonmelanoma skin cancers and melanoma.
SUFU	14	0.003	0.31	The patient has a solitary fibrous tumor.
ERBB4	93	0.06	0.08	This is a man/lady with metastatic colon carcinoma.

Association found in 4 diff. cancers: Bladder Cancer | Head and Neck Carcinoma | Melanoma | Skin Cancer (Non-Melanoma)

ERBB4 is over-expressed in human **colon cancer** and enhances cellular transformation. (Carcinogenesis (2015) doi: 10.1093/carcin/bgv049)

"39 **Notch2** promotes **bladder cancer** progression: Pre-clinical rationale for a novel targeted therapy" (European Urology Supplement 2014; 1569-9056)

"Evidence for differential expression of **Notch** receptors and their ligands in melanocytic nevi and cutaneous malignant **melanoma**" (Modern Pathology (2006) 19, 246–254)

Figure 3: **Example of associations found by our method.** MAF designs the number of occurences of that particular mutation across patients. β is the size effect. The sentence prototype is a randomly chosen sentence from the cluster.

[7] T. Adamusiak and M. Shimoyama, AMIA Summits on Translational Science Proceedings **2014**, 9 (April 2014).

[8] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, J. F. Hurdle and others, Yearb Med Inform **35**, 128 (2008).

[9] P. B. Jensen, L. J. Jensen and S. Brunak, Nature Reviews. Genetics **13**, 395 (June 2012).

[10] M. R. Stratton, P. J. Campbell and P. A. Futreal, Nature **458**, 719 (April 2009).

[11] V. D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, Journal of Statistical Mechanics: Theory and Experiment **2008**, p. P10008 (October 2008), arXiv: 0803.0476.

[12] MSK Researchers Develop Targeted Test for Mutations in Both Rare and Common Cancers.

[13] D. T. Cheng, T. N. Mitchell, A. Zehir, R. H. Shah, R. Benayed, A. Syed, R. Chandramohan, Z. Y. Liu, H. H. Won, S. N. Scott, A. R. Brannon, C. O'Reilly, J. Sadowska, J. Casanova, A. Yannes, J. F. Hechtman, J. Yao, W. Song, D. S. Ross, A. Oultache, S. Dogan, L. Borsu, M. Hameed, K. Nafa, M. E. Arcila, M. Ladanyi and M. F. Berger, The Journal of molecular diagnostics: JMD **17**, 251 (May 2015).

[14] Tumor Sequencing Test Brings Personalized Treatment Options to More Patients.

[15] C. Lippert, J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson and D. Heckerman, Nature Methods **8**, 833 (October 2011).

[16] C. Lippert, F. P. Casale, B. Rakitsch and O. Stegle, bioRxiv (2014).

[17] R. L. Iman and W. J. Conover, Technometrics **21**, 499 (November 1979).

[18] J. M. Bland and D. G. Altman, Bmj **310**, p. 170 (1995).