EXTRACTING SIGNIFICANT SAMPLE-SPECIFIC CANCER MUTATIONS USING THEIR PROTEIN INTERACTIONS

LIVIU BADEA[†]
University Politehnica Bucharest and Bioinformatics Group, ICI
8-10 Averescu Blvd, Bucharest, Romania
Email: badea.liviu@gmail.com

We present a joint analysis method for mutation and gene expression data employing information about proteins that are highly interconnected at the level of protein to protein (pp) interactions, which we apply to the TCGA Acute Myeloid Leukemia (AML) dataset. Given the low incidence of most mutations in virtually all cancer types, as well as the significant inter-patient heterogeneity of the mutation landscape, determining the true causal mutations in each individual patient remains one of the most important challenges for personalized cancer diagnostics and therapy. More automated methods are needed for determining these "driver" mutations in each individual patient. For this purpose, we are exploiting two types of contextual information: (1) the pp interactions of the mutated genes, as well as (2) their potential correlations with gene expression clusters. The use of pp interactions is based on our surprising finding that *most AML mutations tend to affect nontrivial protein to protein interaction cliques*.

1. Introduction and motivation

Although various aspects of the cancer genome, such as gene expression, mutations, DNA copy number changes, or DNA methylation profiles have been studied (mostly) in isolation for more than a decade, their multi-modal, combined analysis has only recently been possible due to large scale projects such as The Cancer Genome Atlas (TCGA), as well as to the dwindling costs of high-throughput sequencing.

Landmark studies of the TCGA have for the first time revealed the genomic changes and their consequences in several cancer types, such as glioblastoma [1,2,3], ovarian [4], breast [5], squamous cell lung cancer [6], colorectal cancer [7] and acute myeloid leukemia [8].

Most of these and other integrated studies of the cancer genome use state of the art methods for analyzing the *separate* data types (such as gene expression, mutation, DNA copy number changes and DNA methylation profiles), and then try to correlate the *separate* findings into a global integrated picture of the cancer genome (for example by searching for mutation enrichment in consensus gene expression clusters, or by comparing miRNA clusters with expression clusters [8]).

Despite numerous attempts at a *joint* analysis of the various data types (as opposed to separate analyses), currently there is no *universally accepted* approach available.

[†] Work partially supported by grant PN-II-ID-PCE-2011-3-0198.

In this paper, we present a joint analysis method for mutation and gene expression data that employs information about proteins that are highly interconnected at the level of protein to protein (pp) interactions, which we apply to the Acute Myeloid Leukemia (AML) dataset obtained by TCGA [8].

Given the low incidence of most mutations in virtually all cancer types, as well as the significant inter-patient heterogeneity of the mutation landscape, determining the true causal mutations in each individual patient remains one of the most important challenges for personalized cancer diagnostics and therapy [18].

For example, since in AML only 3 genes have been found mutated with a frequency above 10% (FLT3, NPM1, and DNMT3A), the state of the art AML study of the TCGA group [8] has used the known gene annotations to determine the genes relevant for pathogenesis (based on a few categories deemed biologically significant by human investigators).

Still, annotations are imperfect and many genes have surprisingly heterogeneous functions. Moreover, annotations reveal nothing about gene interactions (except maybe pathway annotations, which are currently hopelessly incomplete). For example, the NPM1 gene is placed by the TCGA study in a category of its own, solely based on its high mutation rate in AML.

More automated methods are therefore needed for determining the mutations that have caused the disease in each individual patient, the so called "driver" mutations. For this purpose, we are trying to exploit two types of contextual information:

(1) the protein-to-protein (pp) interactions of the mutated genes in question, as well as
(2) their potential correlations with gene expression clusters.

These two types of contextual information are used in a synergistic manner.

The use of pp interactions is based on our surprising finding that *most AML mutations tend to affect complete pp interaction cliques*. More precisely, the protein-to-protein interaction network between AML mutated genes contains a large number of nontrivial maximal cliques (of size ≥ 3).[*]

This is highly surprising given the very low number of somatic mutations in AML, much lower than in all other solid cancers analyzed to date [8]. The fact that mutations tend to affect cliques in the pp interaction network suggests the disruption of biological processes or protein complexes involving the corresponding protein cliques. It is as if such biological processes or complexes can be perturbed by mutations in any of their components. This is important since only very few mutations in AML (or other cancer types for that matter) have an incidence larger than 10%. Grouping mutations based on their pp interactions thereby enhances the statistical power of detecting correlations between mutations (the causal factors) and their transcriptional consequences, such as gene expression subtypes of the disease.

[*] The nontrivial complete maximal cliques of mutated genes have an average size of ~3.

2. Data and preprocessing

2.1. *The TCGA AML dataset*

The TCGA Acute Myeloid Leukemia (AML) dataset was downloaded from the TCGA data portal[†], as well as from the supplementary data of the TCGA landmark publication [8] (in preprocessed form). More specifically, we downloaded:

- *gene expression* data (RNASeqV2 UNC Illumina HiSeq, level 3 RSEM normalized data),
- *copy number variation* data (profiled using Affymetrix SNP6 arrays, level 4 data obtained using Gistic2),
- *somatic mutation* data (obtained using either whole-genome sequencing, or whole-exome sequencing),
- data regarding *gene fusions* (obtained from de novo assembly of RNA-sequencing data), as well as
- *clinical annotations*.

We retained 163 samples with simultaneous gene expression, copy number, mutation, gene fusion and clinical data.

2.2. *Generalized mutations*

Since somatic mutations, copy number aberrations and gene fusions can all act as drivers of the disease in individual patients, we defined "*generalized mutations*" as either:

(1) expressed somatic mutations,
(2) expressed fusion genes, or
(3) significant copy number aberration events.

A somatic mutation in a given gene was considered *expressed* if the expression of the corresponding gene exceeded the expression threshold of 6 (on the log_2 scale).

Since gene fusions have been determined from de novo assembly of RNA-seq data, they were all considered to be expressed.

Copy number aberrations were considered *significant* if

- the corresponding gene's expression levels were not uniformly low (below the expression threshold of 6, mentioned above), and
- they were accompanied by *concordant* gene expression changes with $|Z|>2$ (i.e. amplifications accompanied by gene up-regulation and deletions accompanied by gene down-regulation), and
- the copy number profile had at least a slight correlation (exceeding 0.3) with the gene's expression profile.

There were 2142 genes with generalized mutations in at least one sample, with a total number of 5865 events, of which 1050 expressed mutations and 202 gene fusions (for more details, see

[†] tcga-data.nci.nih.gov

Table 1). Gene fusions g_1-g_2 were recorded as separate generalized mutations in g_1 and g_2 respectively, to allow their mixing with other generalized mutations in those genes.

Table 1. Generalized mutations in 163 AML samples

Generalized mutation type	Number of generalized mutations
CN deletions	3008
CN amplifications	1605
Somatic mutations	1041
Somatic mutations + CN deletions	8
Somatic mutations + CN amplifications	1
Gene fusions	193
Gene fusions + CN deletions	7
Gene fusions + CN amplifications	2
Total	5865

2.3. *Protein-to-protein interaction data*

We used the BioGRID protein interaction database[‡] (version 3.2.101 for Homo Sapiens), which we restricted to the physical interactions. This resulted in 136201 interaction pairs involving 14791 unique human genes.

3. Proteins mutated in AML form pp interaction cliques

Compared to solid cancers, AML genomes have much lower numbers of mutations [8]. This is to be expected, as leukemias do not have to evade the source tissue and metastasize, as solid cancers do. (Along these lines, a two-hit model of leukemogenesis has been proposed by Gilliland [9].)

Interestingly however, restricting the BioGRID pp interaction network to the set of genes mutated in AML, we obtain a large number of pp interaction cliques. More precisely, we obtain 4160 maximal cliques[§] involving the 2142 genes with generalized mutations (of which 3564 *nontrivial* cliques involving more than one gene). The average nontrivial clique size was 2.96. Figure 1 depicts the corresponding distribution of nontrivial maximal clique sizes, showing that mutated genes form many large cliques.

Compared to the complete BioGRID interaction network, the edge density[**] of the network of mutated genes is significantly larger ($2.3 \cdot 10^{-3}$ versus $5.6 \cdot 10^{-4}$), although the average clustering coefficient is smaller (0.1002 versus 0.1758).

[‡] http://thebiogrid.org/

[§] Although the clique decision problem (testing whether a graph contains a clique larger than a given size) is NP-complete, while listing all maximal cliques may require exponential time (as there exist graphs with exponentially many maximal cliques [16]), finding all maximal cliques in our setting is reasonably fast (running times of the order of minutes on a 3GHz machine using a Matlab implementation of the Bron–Kerbosch algorithm [17]).

[**] i.e. the number of edges divided by the maximal possible number of edges, i.e. $n(n\text{-}1)/2$, where n is the number of nodes of the network.

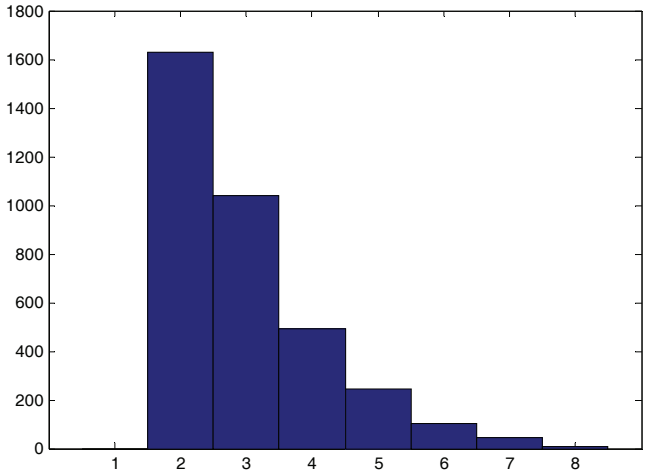

Fig. 1. The distribution of nontrivial maximal clique sizes of mutated proteins in the BioGRID pp interaction network.

Mutations affecting nontrivial protein interaction cliques suggest different ways of perturbing certain key biological processes or protein complexes involving the corresponding cliques. Therefore, although most individual mutations have a low incidence in the AML patient population (thereby masking their possible role in the pathogenesis of the disease), cliques tend to be mutated[††] in a higher number of patients and thus could be used to order mutations in individual patients. The supplementary table '*sample mutations clique cover.xls*' (online at www.ai.ici.ro/ PSB2014/) shows for each patient sample its mutations sorted in descending order of the number of samples in which the largest maximal clique containing the corresponding gene is mutated.

More precisely, in the following, by 'clique' we always mean 'maximal clique'. We denote by M_{ms} the binary mutation matrix (M_{ms}=1 iff sample s has mutation m) and by C_{mc} the clique membership matrix (C_{mc}=1 iff mutated protein m is involved in clique c). We define the cover of a clique c to be the number of samples with mutations in at least a gene m of that clique:

$$\text{clique-cover}(c) = |\ \{\ s\ |\ \exists m.\ M_{ms}=1 \text{ and } C_{mc}=1\ \}\ |.$$

We can also define the largest clique associated to a given mutation m as a clique containing m having the largest clique cover[‡‡]:

$$\text{largest-clique}(m) = c \text{ iff } C_{mc}=1 \text{ and } \forall c' \text{ such that } C_{mc'}=1,\ \text{clique-cover}(c') \le \text{clique-cover}(c).$$

Now, for each sample s, we can sort the mutations m in descending order of the cover of the largest clique associated to m: clique-cover(largest-clique(m)). The top mutations are likely causal,

[††] A clique is said to be mutated in a given sample iff at least one of its genes is mutated in that particular sample.

[‡‡] in case there are several such largest cliques, we arbitrarily choose one.

as they or their interactors are mutated in large numbers of samples. For example, all acute promyelocytic leukemia samples (FAB code 'M3') have the PML and RARA fusion proteins at the top of the list.

4. Joint analysis of gene expression data and mutations using pp interaction data

Although by using protein-to-protein interaction data we have obtained a reasonable ordering of (generalized) mutations w.r.t. their potential causal role in the disease, we still have not made use of all available data to the fullest. For example, we have only employed gene expression data for filtering out mutations in genes that are not expressed, but we have completely ignored any potential similarities in the transcriptomes of samples with different mutations.

In the following, we describe an approach that simultaneously looks for similarities among mutation and gene expression data and, most importantly, is able to extract potentially causal sample-specific mutations, despite their low frequency in the dataset.

By a *direct* joint clustering of gene expression and mutation data, we may only pick up the mutations with the highest incidence. To avoid this, instead of directly clustering mutation data, we cluster the pp interactions of the observed mutations with other mutated proteins. Mutated proteins with similar interactor sets (among the set of mutated proteins) will likely affect the same pathways or protein complexes and produce similar expression changes.

For example, assume sample s_1 is affected by mutation m_1, while sample s_2 is affected by a different mutation, m_2. Even with similar gene expression profiles, s_1 and s_2 may not be grouped into a common cluster k, since we wouldn't know which of the mutations m_1 and m_2 to associate to k. If however, m_1 and m_2 have similar sets of interactors among the other mutated genes $p_1,p_2,p_3,\ldots$, we could *cluster the interactor sets of the mutations* instead of the individual mutations, thereby merging s_1 and s_2 despite their different mutations.

4.1. *The joint clustering of expression and mutation interactor data*

More formally, let s denote samples, g genes, m mutations, k clusters, X_{gs} the gene expression matrix, M_{ms} the binary (generalized) mutations matrix and P_{pm} the binary protein-to-protein interaction matrix involving mutated genes (although the matrix is symmetric, we use distinct p and m indices to distinguish the mutations m from their interactors p).

Now, instead of jointly clustering the gene expression X_{gs} and mutation data M_{ms}, we cluster the gene expression data and the *mutation interactor data* $\sum_m P_{pm} \cdot M_{ms}$:

$$X_{gs} \approx \sum_k G_{gk} \cdot S_{sk} \tag{1}$$

$$\sum_m P_{pm} \cdot M_{ms} \approx \sum_k A_{pk} \cdot S_{sk} \tag{2}$$

where G_{gk}, S_{sk} and A_{pk} are *gene*, *sample* and respectively *mutation interactor cluster matrices*. Note that the sample cluster matrix S_{sk} is common to the gene expression and mutation interactor data factorizations (1) and (2).

Running the nonnegative multirelational decomposition system MNMF[§§] [10,11] with a relative weight w=0.001 for the mutation interactors (to enable the gene expression data to dominate the factorization), we obtain the cluster matrices S_{sk}, G_{gk} and A_{pk} for samples, genes and mutation interactors respectively.

The mutation interactor clusters A_{pk} encode the frequently co-occurring mutation interactors p in the various clusters k, but do not tell us anything directly about the mutations proper. To obtain the *sample-specific mutations* m that lie behind these cluster-specific mutation interactors p, we solve the following *nonnegative least squares problem* (with M'_{ms} as unknown):

$$\sum_k A_{pk} \cdot S_{sk} \approx \sum_m P_{pm} \cdot M'_{ms} \tag{3}$$

using a multiplicative update algorithm that randomly initializes M' and then iteratively applies the following update rule until convergence:

$$M'_{ms} \leftarrow M'_{ms} \frac{(P^T \cdot Y)_{ms}}{(P^T \cdot P \cdot M')_{ms}} \tag{4}$$

where $Y_{ps} = \sum_k A_{pk} \cdot S_{sk}$.

Finally, we can use $MM_{ms}=M'_{ms} \cdot M_{ms}$ as a measure of the significance of mutation m for given clustering. Mutations with higher MM_{ms} are deemed more causally relevant, as they better match the given gene expression clustering. Note that frequently occurring mutations tend to have higher MM scores, especially if they are not at odds with the gene expression clustering.

Figure 2 below is a graphical depiction of the decomposition (1)-(3). The system was implemented in Matlab.

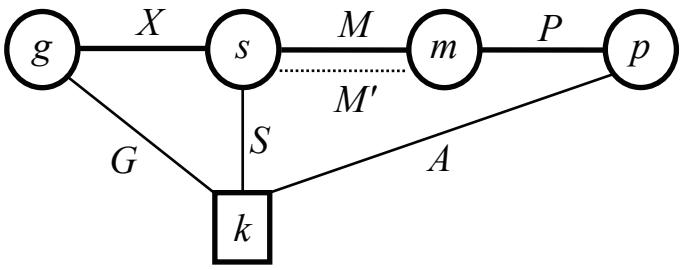

Fig. 2. The relational diagram corresponding to the decomposition (1)-(3). Circles correspond to entities (g genes, s samples, m mutations, p mutation interactors), while the boxed k represents the unknown clusters. Bold edges correspond to the original relations X,M,P, normal edges to inferred entity clusters G,S,A, and the dotted edge to the significant sample-specific mutations M'.

[§§] MNMF is a multirelational generalization of Nonnegative Matrix Factorization (NMF) [13,14] and of simultaneous NMF [15].

4.2. *The dimensionality of the factorization*

Determining the optimal dimensionality n_c of the factorization (1)-(2) is tricky. Similar to Kim and Tidor [12], we performed a series of MNMF runs with progressively larger n_c, ranging from 2 to 50. To avoid overfitting, we performed a similar set of runs on randomized entity matrices and estimated the signal to noise ratio (SNR) as follows:

$$SNR(n) = \frac{\varepsilon_r(n)^2 - \varepsilon(n)^2}{1 - \varepsilon_r(n)^2}$$

where $\varepsilon(n)$ and $\varepsilon_r(n)$ are the relative factorization reconstruction errors for the original and respectively the randomized data. The dimensionality n_c =22 was chosen to maximize the SNR (see Figure 3). Note that the clusters obtained by our nonnegative decompositions should not be confused with partitions of the samples into *disjoint* subgroups. They are rather biclusters corresponding to biological processes that may overlap in the various samples (as well as for certain genes).

We also tried the smaller dimensionality n_c=7 obtained by optimizing NMF consensus sample clustering (a partitional method), as in [8].

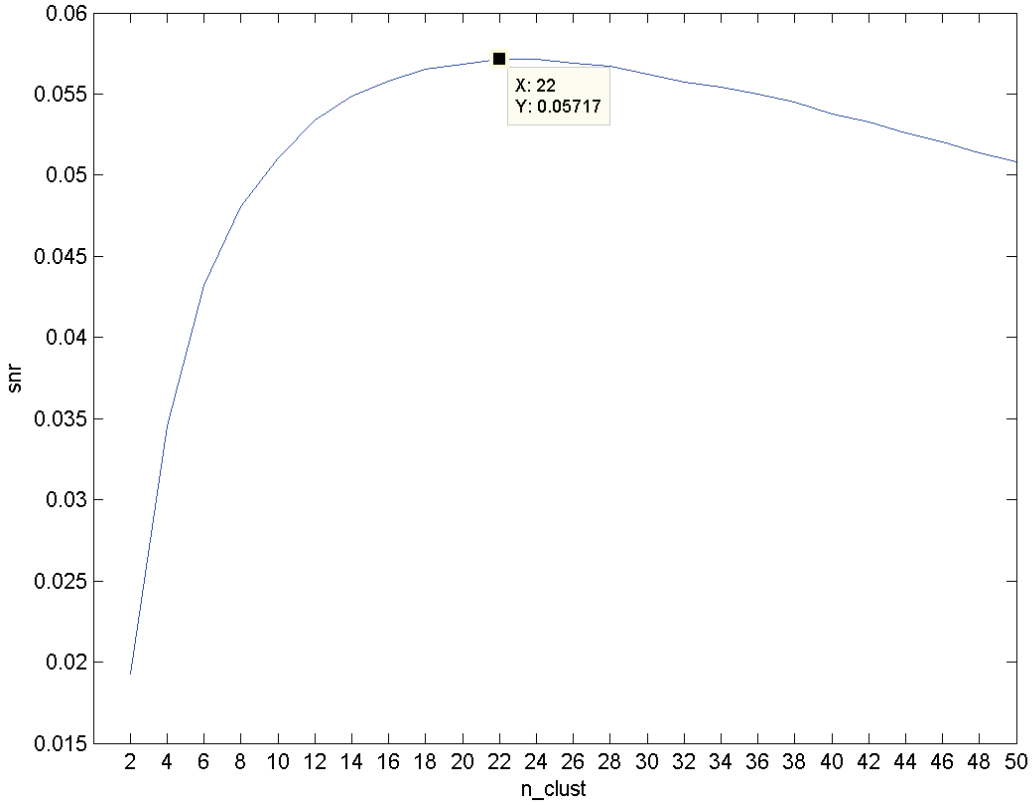

Fig. 3. The estimated SNR for the factorizations ranging from n_c=2 to 50 clusters.

4.3. *Significant sample-specific mutations*

For both n_c=7 and 22, the expression clusters were highly associated (using Fisher's exact test) with the French-American-British (FAB) AML subtypes, as noticed in previous studies (see Table 2 below). In particular, the clustering perfectly distinguishes the Acute Promeylocytic Leukemia (M3) samples from the rest (cluster 3). FAB types M6 and M7 are too weakly represented in our 163 samples (just 1 and respectively 3 samples) to influence the factorization much.

Table 2. Association of clusters with FAB subtypes (n_c=7)

FAB subtype	FAB samples	Best associated cluster (n_c=7)	Cluster samples (n_c=7)	log2(p) (n_c=7)	Best associated cluster (n_c=22)	Cluster samples (n_c=22)	log2(p) (n_c=22)
M0	15	7	33	-8.44	12	23	-24.74
M1	38	5	30	-13.94	14	16	-6.25
M2	39	6	32	-6.06	17	14	-13.18
M3	16	3	16	-41.97	8	16	-41.97
M4	32	1	28	-11.58	22	27	-12.29
M5	17	2	25	-16.36	19	11	-27.13
M6	1	2	25	-2.70	13	14	-3.54
M7	3	7	33	-7.02	13	14	-5.67

The tables '*sample-specific mutations 7 clusters.xls*' and '*sample-specific mutations 22 clusters.xls*' (online at www.ai.ici.ro/PSB2014) list the sample-specific mutations (in descending order of their significance MM_{ms} for each sample *s*) obtained by our approach based on joint clustering of expression and mutation interactor data.

To estimate the concordance of the three mutation significance lists ('*sample mutations clique cover.xls*' from section 3, as well the two tables mentioned in this section), we have computed the average overlap of the top 5 mutations in each sample for all pairs of lists and depicted the results in Figure 4.

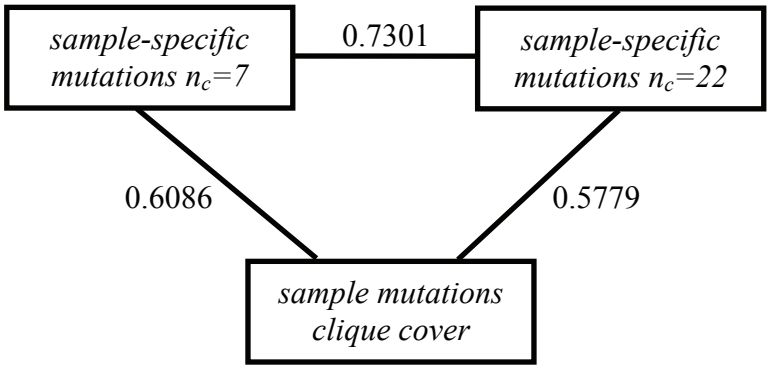

Fig 4. Average overlap between sample-specific mutation significance lists discussed in this paper.

Note that the two sample-specific mutation lists overlap best, as expected, and that the list for n_c=7 is slightly closer to '*sample mutations clique cover*' than the n_c=22 list. The overlap is typically lower for the samples with large numbers of mutations (also as expected).

Careful inspection of the 3 mutation lists shows that we have been able to pick up at least a large fraction (if not most) of the mutations with a causal role in the disease. Virtually all mutations (such as those in NPM1, FLT3, TP53, DNMT3A, etc.), as well as all gene fusions (PML-RARA, MYH11-CBFB, RUNX1-RUNX1T1, etc.) with a known involvement in AML are at the top of the lists of the samples harboring them.

Besides these obvious true positives however, it is difficult to objectively compute accuracy figures for the lists, given the fact that rare individual patient mutations are still largely *terra incognita*. Still, we selected from '*sample-specific mutations 7 clusters.xls*' the sample entries whose top first mutation is *not* among the known AML mutations – we obtained 18 such samples (out of a total of 163), which we list in '*NOT EXPLAINED sample-specific mutations.xls*' (also online). A careful inspection of these samples places them in one of the following 3 categories:

1. Samples with very few detected mutations.
2. A known mutation/fusion is not at the top, but close to it (having significance coefficients close to the top ones).
3. Samples with very many mutations for which known mutations/fusions are far from the top.

    Subcategory 3.1. The first few top entries may contain generalized mutations mentioned in the literature in connection with leukemia.

Obviously, our algorithm does not err too much in categories 1 and 2. Only category 3 (including a few outlier samples with very many mutations) could in principle be improved on – we suspect that they misbehave because those samples do not fit very well in any expression cluster, due to the large numbers of defects accumulated. Table 3 below shows the corresponding samples and their category assignments.

Table 3. Samples with rare mutations

Category	Sample	Comments
1	2946	Only two mutations of unknown role.
1	2995	Only 3 mutations. DDX41(mut) observed by others mutated in AML.
1	3000	Only 3 mutations of unknown role.
1	3008	Only two mutations. Possible role of KAT2B.
2	2832	MLL-MLLT10 fusion close to top.
2	2855	MLLT10-PICALM fusion close to top.
2	2874	IDH2(mut) close to top.
2	2911	MLL-ELL fusion, with MLL significance $3.6 \cdot 10^{-4}$ (close to top significance $5.9 \cdot 10^{-4}$).
2	2940	MLL3(mut) close to top.
2	2955	DNMT3A(mut) with significance 10^{-3} (top $1.2 \cdot 10^{-3}$).

2	3005	MLL-MLLT10 fusion close to top.
3/3.1(?)	2817	CBFB(mut), EZH2(mut), BCR-ABL fusion are far from the top, but LUC7L2(del) at the top (LUC7L2 mutations mentioned in AML).
3.1	2849	MLLT10-PICALM fusion far from top, but at the top, KDM3B (a H3K9 demethylase) is a tumor suppressor linked to leukemia.
3	2882	U2AF1(mut) far from top (significance $1.1 \cdot 10^{-3}$, top $2.3 \cdot 10^{-3}$).
3	2917,2929	KRAS(mut), SETBP1(mut) far from top.
3/3.1(?)	2920	NF1(mut) far from top, but LUC7L2(del) at the top.
3/3.1(?)	2939	MTOR-CDH1 fusion far from top, but LUC7L2(del) at the top.

5. Conclusions

AML, like other cancer types is a heterogeneous disease. But even with multi-genomic data available (related to gene expression, mutations, copy number changes, etc.), finding well-defined *sub-classifications with prognostic and therapeutic value* is still an elusive objective for many cancers (although partial encouraging results have been obtained). This is probably due to the complexity of the biological processes that are perturbed in the disease and which can be affected by a large number of (generalized) mutations. Some of these mutations have a high enough incidence for us to be sure of their causal role in the disease, but many (if not the majority) of the causal genomic events are rare and patient-specific.

In this paper we have shown that we can exploit protein-to-protein interaction data to relate these possibly rare mutations to one another, thereby enabling a better automated detection of the driver mutations in each individual patient. An original feature of our approach is the use of pp interactors of the mutations to enable clustering and especially the back-reconstruction of the significant mutations from the interactor clusters.

HotNet [19], used in the original TCGA publication [8], identified only 4 significantly mutated subnetworks (which are similar to some of our maximal mutation cliques). However, HotNet does not take into consideration the gene expression data, whereas we expect *driver mutations affecting the same pathway* to produce *similar expression changes*.

Future work will address the much more difficult problem of finding *clinically* useful prognostic markers. This will likely require looking at the precise mutations and possibly larger sample sizes, as different mutations in the same pathway or even in the same gene can have significantly different clinical outcomes.

6. Acknowledgments

We are deeply grateful for the invaluable resources put together and made publicly available by the TCGA project. We would also like to thank Andrei Halanay, Daniel Coriu and Jardan Dumitru for discussions, as well as the reviewers for their comments, which helped improve the paper.

References

1. McLendon, Roger, et al. "Comprehensive genomic characterization defines human glioblastoma genes and core pathways." *Nature* 455.7216 (2008): 1061-1068.
2. Noushmehr, Houtan, et al. "Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma." *Cancer cell* 17.5 (2010): 510-522.
3. Verhaak, Roel GW, et al. "Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1." *Cancer cell* 17.1 (2010): 98-110.
4. Bell, D., et al. "Integrated genomic analyses of ovarian carcinoma." *Nature* 474.7353(2011):609-615.
5. Koboldt Daniel C. et al. "Comprehensive molecular portraits of human breast tumours." *Nature* 490.7418 (2012):61-70.
6. Hammerman, Peter S., et al. "Comprehensive genomic characterization of squamous cell lung cancers." *Nature* 489.7417 (2012): 519-525.
7. Muzny, Donna M., et al. "Comprehensive molecular characterization of human colon and rectal cancer." *Nature* 487 (2012): 330-337.
8. Ley, T. J., et al. "Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia." *N. Engl. J. Med* 368.22 (2013): 2059-2074.
9. Gilliland, D. Gary. "Hematologic malignancies." *Current opinion in hematology* 8.4 (2001): 189-191.
10. Badea, Liviu. "Multirelational Consensus Clustering with Nonnegative Decompositions." *Proc. of the 20th European Conference on Artificial Intelligence ECAI 2012* (2012): 97-102.
11. Badea, Liviu. "Unsupervised analysis of leukemia and normal hematopoiesis by joint clustering of gene expression data." *Bioinformatics & Bioengineering (BIBE), 2012 IEEE 12th International Conference on*, (2012): 338-343.
12. Kim, Philip M., and Bruce Tidor. "Subsystem identification through dimensionality reduction of large-scale gene expression data." *Genome research* 13.7 (2003): 1706-1718.
13. Lee, Daniel D., and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." *Nature* 401.6755 (1999): 788-791.
14. Seung, D., and L. Lee. "Algorithms for non-negative matrix factorization." *Advances in neural information processing systems* 13 (2001): 556-562.
15. Badea, Liviu. "Extracting gene expression profiles common to colon and pancreatic adenocarcinoma using simultaneous nonnegative matrix factorization." In *Pacific Symposium on Biocomputing*, vol. 290, pp. 279-290. 2008.
16. Moon, J.W., Moser L. "On cliques in graphs." Israel journal of Mathematics 3.1 (1965): 23-28.
17. Bron, Coen, and Joep Kerbosch. "Algorithm 457: finding all cliques of an undirected graph." *Communications of the ACM* 16.9 (1973): 575-577.
18. Eifert, C., Powers, R.S. "From cancer genomes to oncogenic drivers, tumour dependencies and therapeutic targets." *Nature Reviews Cancer* 12, 572-578 (2012).
19. Vandin, F et al. "Algorithms for detecting significantly mutated pathways in cancer." *Journal of Computational Biology* 18.3 (2011): 507-522.