Poster #35

Subgroups of Gastric Cancer Patients Characterized with an Integrated Large Biomarker Datasets using Association Rules

Williams C[1], Adamczyk B[2], Afshar M[1], Kamali-Moghaddam M[3], Karlsson NG[2], Guergova-Kuras M[1], Lisacek F[4], Mereiter S[5,6], Morley D[1], Parmentier F[1], Polom K[7,8], Roviello F[7], Reis CA[5,6,9,10], Shen Q[3] and Tognetti Y[1]

1. Ariana Pharmaceuticals, Paris, France; **2.** Department of Medical Biochemistry and Cell Biology, Institute of Biomedicine, Sahlgrenska Academy, University of Gothenburg, Sweden; **3.** Department of Immunology, Genetics and Pathology, Science for Life laboratory, Uppsala University, Sweden; **4.** Proteome Informatics Group, SIB Swiss Institute of Bioinformatics, Geneva, Switzerland; **5.** I3S - Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Portugal; **6.** IPATIMUP - Institute of Molecular Pathology and Immunology, University of Porto, Portugal; **7.** Department General Surgery and Surgical Oncology, University of Siena, Italy; **8.** Department of Surgical Oncology, Medical University of Gdansk, Poland; **9.** Faculty of Medicine, University of Porto, Portugal; **10.** Instituto de Ciências Biomédicas Abel Salazar, University of Porto, Portugal

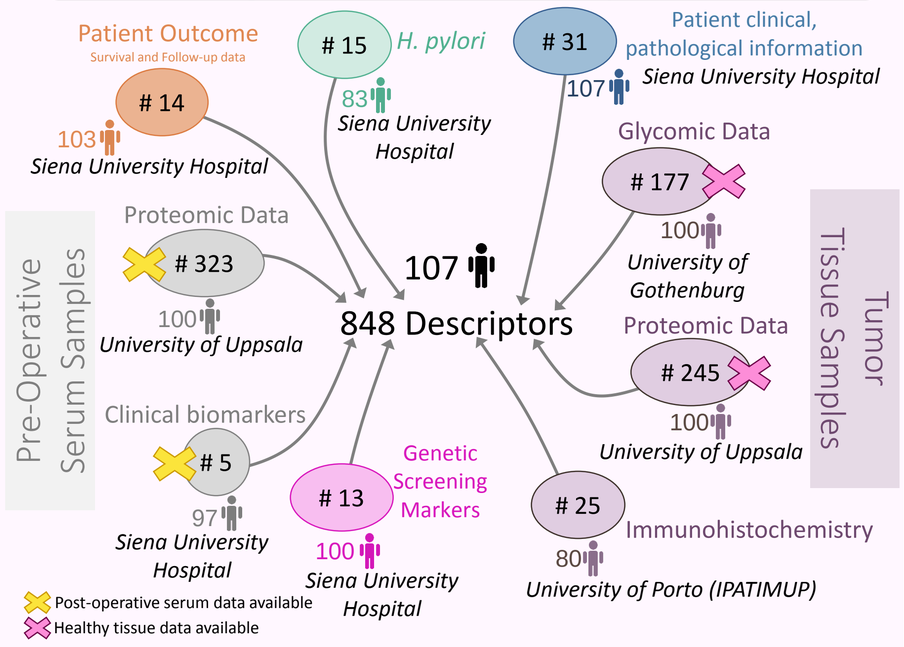
Abstract

Gastric cancer (GC) is one of the most deadly form of cancer worldwide, partly due to the lack of early diagnosis[1,2]. Availability of molecular data, characterizing cancer patients and their tumour, is required for improved diagnosis and prognosis of patients. The commitment of clinicians to provide a precision medicine approach in the diagnosis, prognosis and treatment of GC drives the need for better biological markers. We describe a retrospective study collecting glycomic, proteomic, immunohistochemistry, *Helicobacter pylori*, and blood biomarker measurements from tissue and serum samples of 107 gastric cancer patients that underwent surgery in the Division of Surgical Oncology, at Tertiary University Hospital of Siena, Italy. In this work, we developed a specific framework dedicated to the integration of multiple datasets from several heterogeneous sources and platforms. Experimental data was integrated with clinical, historical and survival information available for patients providing a large heterogeneous database of 848 variables. This study identified subgroups of patients of clinical importance using a Machine Learning methodology (KEM®, Knowledge Extraction and Management[3]) that provides, through exhaustive exploration of all relationships between patient's variables, an hypothesis-driven approach helping interpret this broad database and thus identify actionable hypotheses. We systematically extracted all logical associations between experimental measures and clinical outcomes obtaining a knowledge base of over 1000 associations identifying potential disease risk markers.

Stomach Cancer | 107 patient cohort | 848 variables | Association Rules | KEM® Framework | Hypothesis-driven | Patient characterization | Disease risk markers

1 • Data Integration

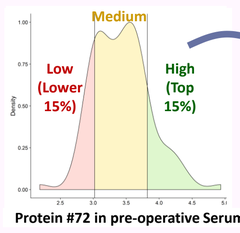
a • Integration of Data from Multiple Platforms

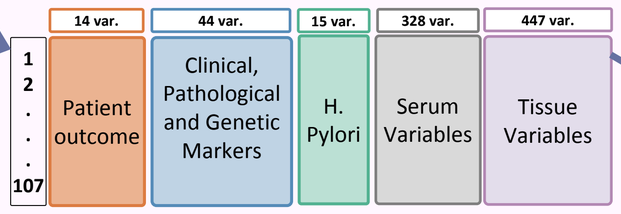


b • Data Transformation: *from Continuous to Discrete*

Protein #72	Protein #72 discretized
4,30	High
4,94	High
3,04	Medium
...	...
3,77	Medium
2,97	Low
2,93	Low

The experimental data was discretized and imported into the KEM® framework. Proteomic and glycomic data in serum and tissue samples were discretized into three bins: High bin (top 15%), Medium bin and Low bin (lower 15%). Patient outcome, clinical, pathological, genetic markers, h. *pylori* and immunohistochemistry variables were already discrete.





2 • Analysis Steps

a • KEM® Framework

Associations Rules: definition, metrics

KEM® generates association rules $\mathbf{Var_i \rightarrow Var_j}$ in an exhaustive manner. These rules are characterized by 4 metrics that help ranking them.

	Var 1	Var 2	Var 3 (Endpoint)
Patient 1	low		True
Patient 2	low		True
Patient 3	low		True

Rule: Var 1 = low → Var 3 = True

Metrics: Support, Confidence, Lift, P-value

Support number of times that the rule is checked in the dataset

Confidence proportion of cases verifying Var1 = low and Var3 = True.

Lift ratio of the observed support to that expected if Var1 = low and Var3 = True were independent.

P-value Fisher exact test

Generations of all relationships between all variables in discrete database

KEM® Tools

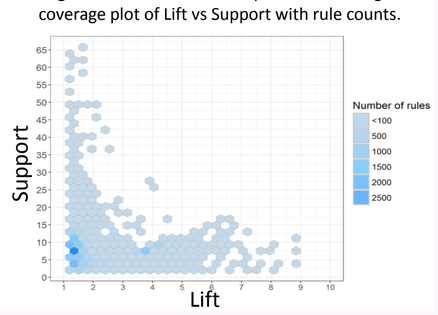
1) Exhaustive rule extraction with KEM®

2) Visualisation and Exploration of subset of best rule with KEM® Query

b • Filtering Strategies to Explore Rules

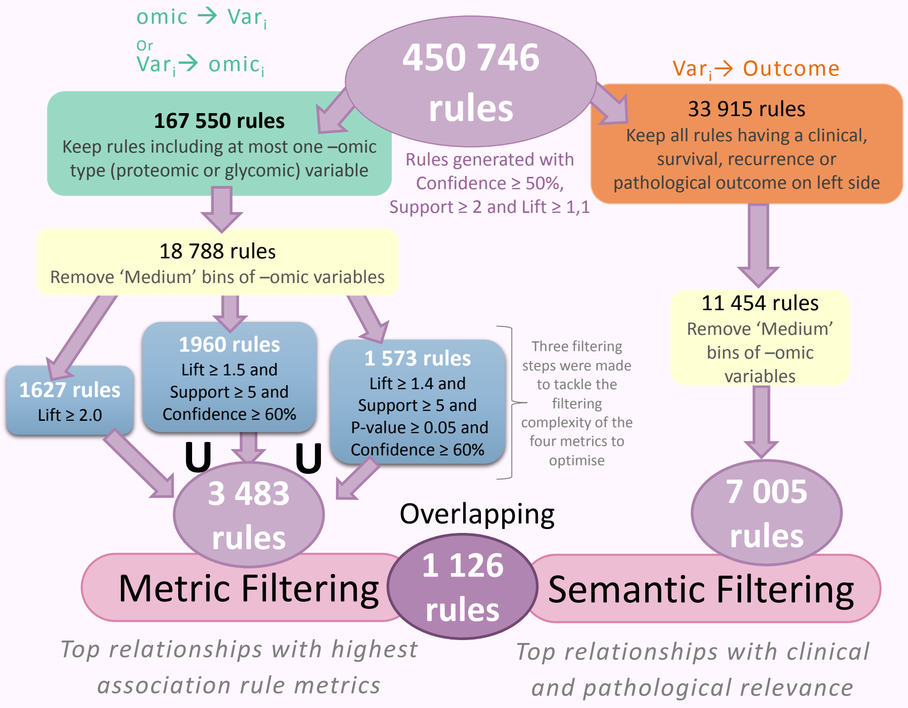
Rules metrics assessment & filtering

Figure 1: Metric threshold optimisation using a coverage plot of Lift vs Support with rule counts.



Exploration of the most interesting and strongest generated associations was undertaken by applying two filtering strategies:

- First strategy optimises rules metrics in a four dimensional space. Coverage plots were used to find the threshold values for each metric (example in Figure 1).
- Second strategy keeps all rules in which a right variable describes a clinical, pathological or survival outcome.



3 • Results

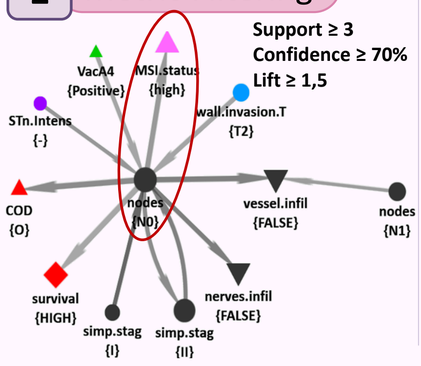
LYMPHNODE METASTASIS

MSI | MSS | Molecular Subtypes

Lymphnode metastasis is considered as one of the most important prognostic predictors in gastric cancer.

→ **VALIDATION:** *Confirming previously described relationships*

1 Metric Filtering



2 Semantic Filtering

The rule of interest found by filtering by metrics (circled here on left-hand side) was found in a predictive direction (Var→ Outcome) in the semantic filtering subset:

MSI.status { High } → Nodes { N0 }

Left	Right	Support	Conf.	Lift	Pvalue	#Left	#Right
MSI status (HIGH)	Nodes (N0)	20	53%	2.07	3,05E-06	38	27

- Microsatellite instability has been defined as distinct molecular subgroup[4] linked to fewer lymphnode metastasis and overall improved prognosis[5].
- The results are aligned with this description showing patients with high MSI are associated with N0/N1 status.

Variable code (names shown in KEM®)	Description	Variable Category (seen in KEM® Query in brackets)
nodes	Number of lymphnode metastasis	**N0, N1, N2, N3**
wall.invasion.T	Tumour growth in stomach walls	**T1, T2, T3, T4**
MSI.status	Microsatellite Status	**High, Stable**
VacA4	VacA4 allele of Helicobacter Pylori (measured using PCR[6])	**Positive, Negative**
STn.Intens	Sialyl Tn Intensity	**-, x, xx, xxx**
COD	Cause Of Death	**O** (other), **GC** (Gastric Cancer)
survival	Months of survival after first operation	**LOW** (≤ 2years), **MEDIUM, HIGH**(> 5years)
nerves.infil	Nerves infiltration	**TRUE, FALSE**
vessel.infil	Vessel infiltration	**TRUE, FALSE**

SIMPLIFIED STAGING

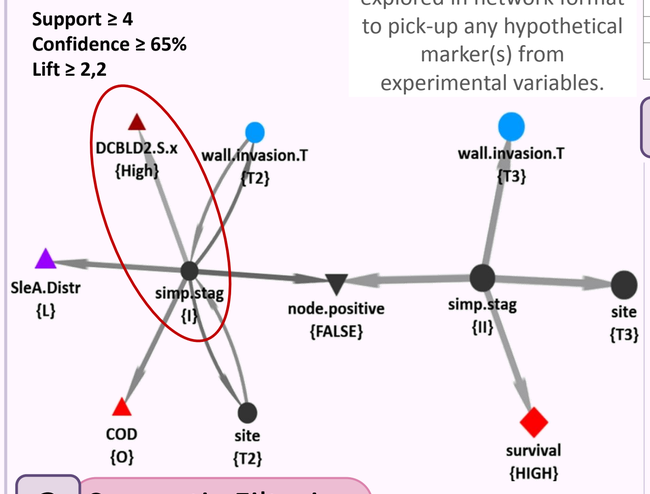
Simplified Staging is a concise description of TNM (Tumour, Nodes and Metastasis) staging of gastric cancer patients which helps to predict the progression of the disease.

→ **DISCOVERY:** *Uncovering new relationships* | TNM Staging | Serum Marker

Simplified Staging Categories	TNM staging description	No of patients
I	IA, IB	9
II	IIA, IIB	24
III	IIIA, IIIB, IIIC	57
IV	IV	17

1 Metric Filtering

In a first step, associations filtered by metrics were explored in network format to pick-up any hypothetical marker(s) from experimental variables.



Variable code (names shown in KEM®)	Description	Variable Category (seen in KEM® Query in brackets)
Site	Tumour Site	**D, I, II, III**
DCBLD2.S.x	DCBLD2 in pre-op serum	**High, Low**
SleA.Distr	Sialyl-Lewis A distribution	**L, M, H**

2 Semantic Filtering

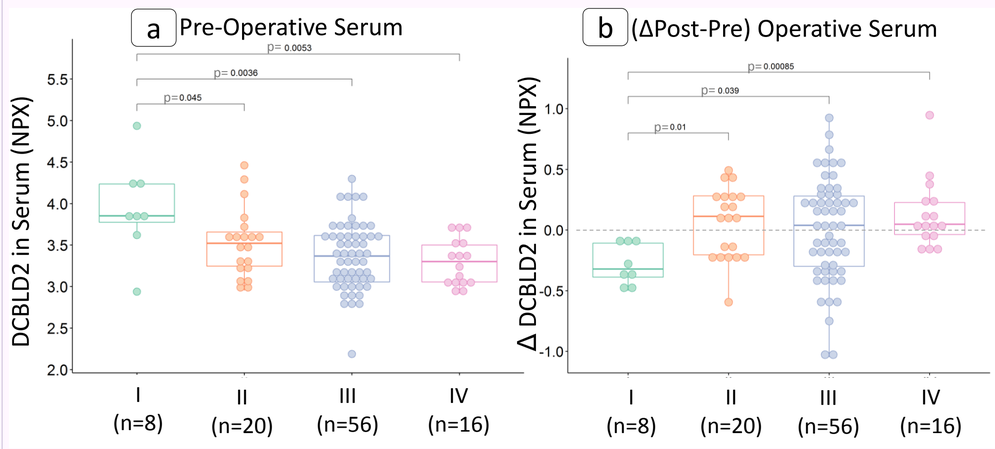
In a second step the rule of interest, circled above, was checked if it was present in a predictive direction (Var→ Outcome) in the semantic filtered rules subset:

DCBLD2.S.x { High } → Simp.Stag { I }

Left	Right	Support	Conf.	Lift	Pvalue	#Left	#Right
DCBLD2 in pre-op serum (HIGH)	Simp Staging (I)	11	73%	1,4	0,079	15	57

3

The protein shows a significant difference according to stage in serum samples (Figure a), with higher NPX values linked to Stage I group. Protein NPX levels in Stage I group significantly decreased in serum between post and pre operation compared to other staging groups (Figure b).



The protein expression in tumour tissue had no significant difference across simplified stages.

Conclusion

KEM® platform helps generate new hypotheses and validate previous knowledge from associations. This work describes a data-driven framework using association rules to extract knowledge from an integrated database. A subset of identified relationships were presented and discussed.

This work demonstrates the potential of combining powerful Machine Learning tools, experimental glycomic, proteomic data and clinical information to discover potential markers for non-invasive diagnosis and prognosis of gastric cancer.

References

1. World Health Organization. (2017) Cancer: Fact Sheet No 297. WHO. http://www.who.int/mediacentre/factsheets/fs297/en/.
2. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, and Jemal A. (2012) Global cancer statistics. CA Cancer J Clin. 2015;64:87–108.
3. Afshar M, Lanoue A, and Sallantin J. (2007) Multiobjective/Multicriteria Optimization and Decision Support in Drug Discovery. Comprehensive Medicinal Chemistry II. Volume 4, edn. 2007: 767-774.
4. Cancer Genome Atlas Research, N. (2014) Comprehensive molecular characterization of gastric adenocarcinoma. Nature. 513, 202-209
5. Zhu L, Li Z, Wang Y, Zhang C, Liu Y, and Qu X. (2015) Microsatellite instability and survival in gastric cancer: A systematic review and meta-analysis. Molecular and Clinical Oncology, 3(3), 699–705.
6. Marrelli D, Pedrazzani C, Berardi A, Corso G, Neri A, Garosi L, Vindigni C, Santucci A, Figura N, and Roviello F. (2009) Negative Helicobacter pylori status is associated with poor prognosis in patients with gastric cancer. Cancer. 2009;115:2071–2080.

Funding: We acknowledge the support from the European Union, 7th Framework Programme, Gastric Glyco Explorer initial training network: grant no. 316929.

www.arianapharma.com Email: m.afshar@arianapharma.com