Simultaneous Search of Genomic and Proteomic Biomarkers in Human Colorectal Cancer

García Amaia[1], Freije Ana[1], Armañanzas Rubén[2], Inza Iñaki[2], Ispizua Ziortza[1], Heredia Pedro[1], Larrañaga Pedro[1], López Vivanco Guillermo[3], Suárez Tatiana[1], Betanzos Mónica[1]

[1.]Department of Biotechnology, GAIKER Technological Centre, Parque Tecnológico, Edificio 202, 48170 Zamudio, Bizkaia, Spain.
[2.]Department of Computer Science and Artificial Intelligence, University of the Basque Country, P.O Box 649, E-20080 Donostia-San Sebastián, Spain
[3.]Service of Medical Oncology, Plaza de Cruces s/n, 48903, Barakaldo, Bizkaia, Spain.

ABSTRACT

A simultaneous study for searching genomic and proteomic biomarkers is being carried out in human colorectal samples. A total of 133 samples, 60 colorectal tumor samples, 60 paired non tumoral samples corresponding to different stages of the disease, and 13 control non cancerous samples were collected and analyzed by genomic and proteomic approaches.

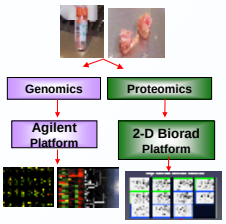


Fig. 2. Diagram of the experimental approach.

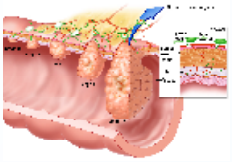


Fig. 1. Progression of colorectal cancer

1. INTRODUCTION

Colorectal cancer (CRC) is the second cause of cancer death in western countries. The success of the therapy depends on an early diagnostic, the knowledge of the biological behaviour in each tumor, and its susceptibility to drugs. DNA microarray technology allows the measure of the mRNA expression level of thousands of genes simultaneously. Proteomic approach based on 2D-SDS PAGE strategy permit to identify changes in protein expression induced by cancer involved processes, and to identify protein biomarkers.

2. MATERIALS AND METHODS

Tissues and patients. A total of 133 tissue samples (120 from patients with CRC in different stages, and 13 samples from patients with no colorectal cancer) were obtained from Cruces Hospital (BIOEF). The 120 samples consisted in 60 tumour samples and 60 paired non tumoral samples.

Patient	TNM	DUKES	Age	Sex	Patient	TNM	DUKES	Age	Sex
M1	IV	D	46	F	M2	IV	D	46	F
M11	IIIA	C1	87	F	M12	IIIA	C1	87	F
M14	IIIC	C2	68	F	M13	IIIC	C2	68	F
M20	IV	D	77	F	M19	IV	D	77	F
M25	IIA	B2	71	F	M26	IIA	B2	71	F
M35	IIIB	C2	73	F	M36	IIIB	C2	73	F
M41	IIA	B2	66	F	M42	IIA	B2	66	F
M55	IIIB	C3	67	F	M56	IIIB	C3	67	F
M65	IIIB	C2	71	F	M66	IIIB	C2	71	F
M88	IIA	B2	76	F	M87	IIA	B2	76	F
M90	IIIB	C3	57	F	M89	IIIB	C3	57	F
M106	IIA	B2	75	F	M105	IIA	B2	75	F
M112	IIIA	C1	62	F	M109	IV	D	69	F
M116	IV	D	72	F	M111	IIIA	C1	62	F
M5	IIA	B2	69	M	M6	IIA	B2	69	M
M8	IIB	B3	68	M	M7	IIB	B3	68	M
M10	IV	D	63	M	M15	IV	D	81	M
M22	IA	B1	73	M	M16	IV	D	81	M
M23	IIIC	C3	47	M	M21	IA	B1	73	M
M29	IIA	B2	72	M	M24	IIIC	C3	47	M
M37	IIB	B3	73	M	M40	IIA	B2	55	M
M39	IIA	B2	55	M	M44	IA	B1	77	M
M43	IA	B1	77	M	M50	IA	B1	46	M
M49	IA	B1	46	M	M54	IV	D	50	M
M53	IIIC	C2	47	M	M59	IV	D	83	M
M53	IV	D	50	M	M63	IV	D	74	M
M60	IV	D	83	M	M69	IIIB	C2	65	M
M64	IV	D	74	M	M72	IIA	B2	57	M
M70	IIIB	C2	65	M	M81	IIA	B2	71	M
M71	IIA	B2	57	M	M95	IIA	B2	63	M
M74	IA	B1	71	M	M100	IA	B1	76	M
M78	IA	B1	68	M	M119	IIIB	C2	67	M
M80	IA	B1	59	M	M9	IV	D	63	M
M83	IIIC	C2	70	M					
M94	IIA	B2	63	M					
M104	IIA	B2	78	M					
M107	IIA	B2	60	M					
M120	IIIB	C2	67	M					

Table 1. Clinical and pathologic data of patient tumor (rigth) and non-tumor (left).

2.1 GENOMICS

Tissue samples were preserved in *RNA later Stabilization Solution* (Qiagen) and stored at -80ºC.

RNA extraction. Total RNA was extracted from all the samples using the *RNAeasy Mini Kit (Qiagen)*. RNA quality and quantity was determined with the *Agilent 2100 Bioanalyzer (Agilent Technologies)*. We used the RIN algorithm (RNA Integrity Number, *Agilent Technologies*) as a quality standard to select the samples. We synthesized and labelled the cRNA using the *Agilent Low RNA Input Fluorescent Linear Amplification Kit (Agilent Technologies)*

cRNA hybridization. The selected samples were hybridized onto the *Agilent Human 1A 60-mer oligo microarrays (Agilent Technologies)* and the microarrays were scanned using the *GenePix 4000B Scan (Axon Instruments)*. Images were analyzed with Genepix 6.0 (Axon Instruments) and data were filtered and normalized with Acuity 4.1 (Axon Instruments).

Experimental Design. Tumoral and non-tumoral samples were hybridized against a common RNA control pool. As none of the control non-cancerous samples presented an acceptable RNA quality they were discarded and we collected RNA from paired non-tumoral samples to form the NT Pool. The tumoral samples were labelled using Cy-5 dye (red) and the "pool" was labelled with Cy-3 dye (green)[1].

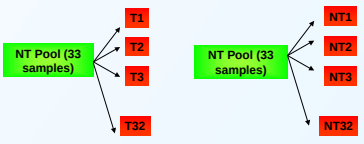


Fig. 3. Hybridization design

Spot quality metrics. Reliability in the microarray probes are tackled by applying three different widely used quality metrics[2]: *fluorescent intensity measurement* quality, *background flatness* quality and *signal intensity consistency* quality. In basis of these three metrics a global quality metric with values between 0 and 1 is computed for each spot in each microarray.

Imputation of lost values. Collateral undesirable problems, such as small fibres inside the array, or an incomplete hybridization, can cause a spot value to be lost. In order to complete all these lost values we used the *KNNImpute*[3] procedure which has been proven as one of the best imputation techniques in the microarray domain.

Intraclass ratio differences. It is not expectable to find big differences between the expression ratios of a gene in between the same type of tissue. But, due to the heterogeneity of the cells included in the biopsies, genes with expression differences bigger than *2-fold* in the same kind of tissue are discarded.

Global machine learning approach. On the basis of the CRC state of each patient, we propose a supervised classification problem, or class prediction problem. The classification dataset is then composed of 64 instances from four different classes with cardinalities: 33 non-tumour, 13 Dukes B, 10 Dukes C and 8 Dukes D.

Discretisation policy. To apply the following statistical techniques the continuous expression values have to be discretise. Attending to the expected biological behaviour -under, baseline or over expressed-, the values are discretised using an *Equal Width* policy with three intervals.

Univariate statistical metrics. Using the supervised approach we can univariatelly measure the relevance of each gene (from now on called *variable*) in the problem. Six different statistical metrics[4] were computed: *Mutual information*, *Euclidean distance*, two versions of the *Kullback-Leibler divergence*, *Matusita* and *Battacharyya* metrics. Sorting the variables by means of their coefficients, we can construct six different importance rankings.

Consensus univariate relevance. Individually, the univariate relevance metrics may be biased owing to the low number of instances. In these scenarios and to achieve a more dependable result it is better to put all the metrics together into a consensus. The consensus among the six original rankings is made up using the average position of each variable over all the rankings. The final consensus ranking shows the statistical univariate relevance of each probe in the supervised problem.

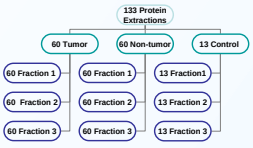


Fig.6. Schematic scheme of the samples processed.133 Sequential Protein Extractions were made.60 paired samples tumor-nontumor and 13 control samples were processed obtaining 399 protein fractions.

2.2 PROTEOMICS

Protein Extraction and Quantification. Sequential protein extraction was made by using ReadyPrep sequential extraction kit (Bio-Rad) based on the differential solubility of the proteins [5] Three fractions were obtained, and the soluble fraction (Fraction 1) was cleaned up with the Ready Prep 2-D Cleanup (Bio-Rad).

Protein concentration was determined using the RC-DC protein assay (Bio-Rad), and EZQ Protein Quantitation Kit (Molecular Probes).

IEF assays. IEF assays were made using Bio-Rad pH 4-7 immobilized pH gradient (IPG) strips to separate proteins according to their isoelectric points. The IPG strips were loaded with 40 µg of protein sample, subjected to active rehydration, and focused using PROTEAN IEF Cell (Bio-Rad) for a total of 35kVh.

2D-SDS assays. IPG strips were loaded and on Criterion gels 2D Precast (Bio-Rad) 8-16% acrilamide. Electrophoresis was carried out for 90 minutes on a Dodeca Criterion Cell at 200V. The gels were stained with SYPRO Ruby dye (Bio-Rad). The gel images were captured with VersaDoc (Bio-Rad) and analyzed using The PDQuest Software. Protein identification was made by MALDI-TOF analysis.

3. RESULTS

3.1 GENOMICS

After the quality analysis of all the RNA samples and based on the electropherograms and the RIN number values obtained we decided to discard all those samples with a RIN below 6. Consequently, we selected a total of 32 tumoral and 33 non-tumoral samples for the microarray gene expression analysis.

After scanning the 65 microarrays we removed the "control spots" and normalized the data obtained by Lowess Normalization [6].

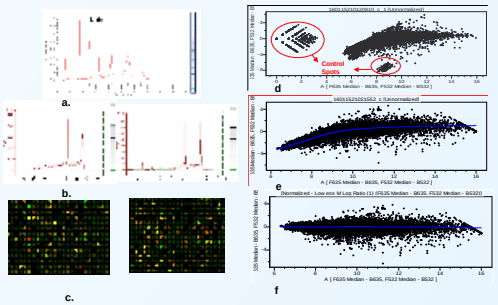


Fig 5. A) Electropherogram of the RNA 6000 Ladder (Ambion). B) Two samples of tumor (left) and paired non tumoral sample (right). C) Microarray image representing the hybridization of each sample above with the pool. D) Data obtained form one microarray showing the "control spots". E) Unnormalized data after removing the "control spots". F) Data normalized by Lowess.

Once the control spots were removed from the data, the total number of probes descended from 22,574 to 17,986 probes. On the quality metrics filter process the acceptance threshold was set up in an average of 0.99 quality value; a total of 11,120 probes surpassed this stage. The imputation algorithm was run with a K value of 15 neighbors. From the 722,800 number of total spots, there were only 1,04% of lost values (7,534 probes) to imputate. The last filtering step removes 3,016 probes that showed differences bigger than *2-fold* in between each of the four classes of tissues. A total of 8,104 probes composed the final dataset.

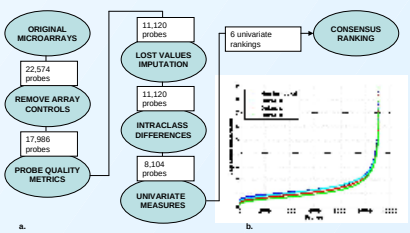


Fig 6. A) Overall process of the data analysis, for each stage the number of probes that surpass the stages are included in the boxes. B) Intraclass diperssion measure for the 11,120 filtered quality probes.

Rank	Gene	Description
1	ENC1	Ecodermal neural cortex 1- Altered expression may contribute to brain tumour development and CRC. Marker of neuronal maturation.
2	ACAT1	Ac-Coa acetyltransferase 1.Mutations in the corresponding gene are associated with 3-ketothiolase deficiency.
3	FLJ20539	Protein of unknown function, has low similarity to uncharacterized human KIAA1906.
4	SNRPB2	Small nuclear ribonucleoprotein polypeptide B. Functions as an autoimmune antigen in systemic lupus erythematosus (SLE) and other rheumatic diseases patients
5	TERA	Protein of unknown function, has high similarity to uncharacterized mouse Tera.
6	TCF3	Transcription factor 3. HLH transcription factor regulates immunoglobulin gene expression; chromosomal rearrangements leading to the expression of a E2A - PBX1 chimeric protein are associated with acute leukemias.
7	I_948907	Member of the DEAD or DEAH box ATP-dependent RNA helicase family. Moderate similarity to ATP-dependent RNA helicase (*S. cerevisiae* SPB4), which is required for processing of 25S ribosomal RNA precursor.
8	MCT-1	Multiple copies in T-cell malignancy. Putative oncogene is involved in cell cycle regulation and participates in positive control of cellular proliferation through the regulation of CDK activity, amplified and overexpressed in T-cell lymphomas.
9	ACO2	Aconitase 2 mitochondrial (aconitate hydratase), catalyzes the conversion of citrate to cis-aconitate in the tricarboxylic acid cycle, may be involved in iron homeostasis; deficiency may be associated with lifelong exercise intolerance.
10	PMAIP1	Phorbol-12-myristate-13-acetate(PMA)-induced protein 1, a likely immediate early response gene; highly expressed in adult T-cell leukemia cells.

Table 2. First ten genes in the consensus relevance ranking.

3.2 PROTEOMICS

Protein quantification was estimated before and after clean up step. Rehydration sample buffer was found to interfere with both Bradford, and RC-DC (BioRad) methods. Therefore, we assayed a novel method based on fluorescence (EZQ Protein Quantitation Kit, Molecular Probes), and we found that this method does not interfere with any sample buffer used in our experiments (Figure 7).

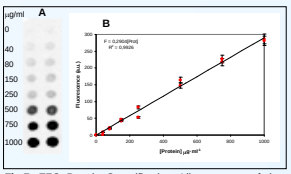


Fig.7. EZQ Protein Quantification. Linear range of the assay for BSA , 0 to 1 mg/ml .The fluorescence was read with the VersaDoc (Bio-Rad) **(A),** and Perkin Elmer LS-50B spectrofluorometer **(B)**.

Proteomic analysis of Fraction 1 is being performed by 2 dimensional gels. In figure 8 a proteome analysis example of a colorectal cancer patient (stage C1) is showed. The assays were made by triplicate, and a matchset containing all the spots from each gel was generated and analyzed using PDQuest software. In this case, a total of 283 spots were found, with variability among gels of 19.08% (229 matched spots, and 54 unmatched spots). 6 spots were selected, picked and identified by MALDI TOF.

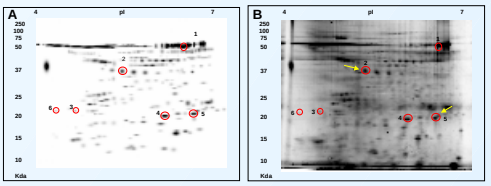
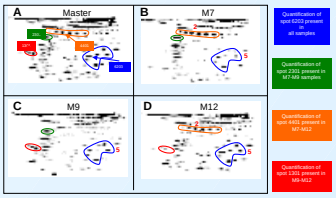


Fig.8.Proteome analysis of patient 12. **A.** Matchset generated from 3 gels from the same patient. **B.** Filtered image of one of the three gels obtained from patient 12, stained with SYPRO Ruby. Spots indicated in red circle were submitted for identification by MALDI-TOF. The identified spots are as follow: Spot 2: Haptoglobin; Spot Peroxiredoxin-2. The identification of the remaining spots was not conclusive.

A proteomic analysis of 3 patients of different colorectal cancer stages (B3; C1; and D) (see Table 1) is showed in figure 9. An High level matchset was created from the individual matchsets obtained from each patient gels as described above. The spot count gave a total of 298 spots in the high level matchset. Difference among cancer stages was 33.89%, indicating differences in proteome related to the progression of the disease.

Using an intermediate stage (C1) as a reference, we determined the unmatched spots ratio in other cancer stages (D, and B3). Stage D showed a 34% difference of unmatched spots, while stage B3 showed a 21.33% of unmatched spots.



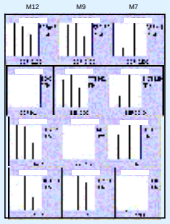


Fig.9. Identification of proteins in matchsets obtained from three 2D gels from three different tumoral samples (M7,M9 and M12). Colours indicate protein patterns common to all (blue), M7-M9 (green), M7-M12 (orange) and M9-M12 (red) samples. Numbers refer to the SSP (Standard Spot Number) assigned automatically in the master matchset.
A. High Level Matchset obtained from merging all the proteins observed in matchsets from M7, M9 and M12 tumoral samples in different cancer stages. **B.**Matchset of tumoral sample in carcer stage B (150 spots), **C.** Cancer stage C (214 spots), **D.** cancer stage D (202 spots). **E.** Diagram of spot quantities.

Spot	Accession number	Description
2	P00738	**Haptoglobin.** Combines with free plasma hemoglobin, preventing loss of iron through the kidneys and protecting the kidneys from damage by hemoglobin. A haptoglobin -like protein that may be secreted by colon cancer cells shows promise as a serum marker for the presence of
5	P32119	**Peroxiredoxin-2(Natural killer cell -enhancing factor B).** Involved in redox regulation of the cell. Reduces peroxides with reducing equivalents provided through the thioredoxin system. Might participate in the signaling cascades of growth factors and tumo r necrosis factor-alpha by regulating the intracellular concentrations of H_2O_2[8]

Table 3. Mass Spectroscopy results of spots 2 and 5.

CONCLUSIONS

We have already obtained a tentative model (first ten genes are showed in table 2) for the classification of cancerous and non-cancerous samples based on their gene expression profile. We are now in the validation process of this model, and it is of the outmost importance for us to check its potentiality for diagnosis/prognosis.

From the machine learning point of view we envision the building of different classification models. Furthermore, the search for statistical reliable dependencies could bring us some light regarding the complex nature of human CRC. An interesting approach would be as well to try to look for the inherent relationships between the genomics and the proteomics analysis.

Our preliminary results from proteomic studies suggest that differences in protein expression could be related to differential stages of disease. From the identified proteins, **Haptoglobin** was found to be present in both patient M7 (stage B) and M12 (stage C), but was absent in patient M9 (stage D), while **Peroxiredoxin-2** was present in all samples of the 3 stages.

Bibliografía

[1] van't Veer L.J et al. *Nature*, 2002, 415:530-535
[2] Chen Y. et al. *Bioinformatics*, 2002, 18(9), 1207-1215
[3] Troyanskaya O. et al. *Bioinformatics*, 2001, 17(6), 520-525
[4] Molloy M.P. et al. *Electrophoresis*,1998,19(5),837-44.
[5] Quackenbush J. *Nature Genetics*, 2002, 32:496-501
[6] Ben-Bassat M. Handbook of Statistics, 1982, 2:773-791
[7] Bresalier RS.et al.*Gastroenterology*,2004,127(3):701-8
[8] Zhang P. *J. Biol Chem*. 1997, 272(49): 30615-30618