Project Acronym	Mut2Dis
Project Code	PIOF-GA-2009-237225
Project Title	New methods to evaluate the impact of single point protein mutation on human health.
Periodic Report	Returning Phase, Sep 2011 – Aug 2012 (12 months)

WORK PROGRESS AND ACHIEVEMENTS DURING THE PERIOD

1. Progress towards objectives and details for each task

1.1 Development of disease-specific predictors

For the accomplishment of the 5th task, EC started in the last part of the outgoing phase, collecting cancer-related missense Single Nucleotide Variants (mSNVs) selecting only mutations with disease names associated to the MESH term *"neoplasm"*. During the returning phase the previous set was compared with a manually curated dataset of cancer driver mutations to select a set of cancer-causing mutations and remove possible passenger cancer mSNVs not directly cause of the pathological state. After this procedure, we collect a set of 3,163 cancer-causing mSNVs from 74 proteins. For training and testing the cancer specific predictor, we selected two different set of non cancer-causing mutation. The first subset composed of polymorphisms from the SvissVar dataset with frequency higher than 0.01 and chromosome sample count higher than 49 in the dbSNP database build 131. The second subset of negative mSNVs by the 50% of the polymorphisms in the first subset and by 50% of disease-related variants in SwissVar, which are not associated to the MeSH term "*neoplasms*". Comparing the distributions of the features for positive and negative subsets, we select the most discriminative ones for implementing our Support Vector Machine (SVM) to predict cancer-causing and non cancer-causing mSNVs.

The SVM classifier (SPF-Cancer) takes in input a 52 elements vector encoding for protein sequence and functional information. The composition of the input vector is summarized in Fig. 1.

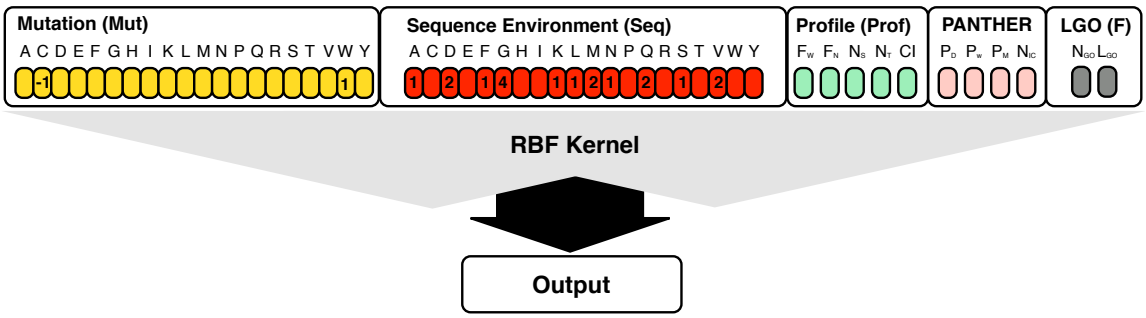

Fig 1. Input features of the cancer-specific SVM-based method (SPF-Cancer). The input feature encodes information about the mutation (yellow), the sequence environment around the mutation (red), the sequence profile (green), PANTHER output (pink) and protein function (black).

The SPF-Cancer has been trained and tested using a 20-fold cross-validation procedure on the CNO dataset, which composed by 3,163 cancer causing mSNVs and 3,163 randomly selected polymorphisms from SwissVar. The cancer-specific method has been tested on the CND dataset, that is composed by 3,163 cancer causing mSNVs and 1,581 randomly selected polymorphisms and 1,582 non-cancer disease-related mSNVs from SwissVar. In addition, using the annotation reported in SwissVar, we split the dataset in the subset related different types of cancer. For the purpose of the comparison of our method and previously developed ones we collected a Synthetic dataset including a set of previously *in silico* generated passenger mSNVs. The composition of each dataset is reported in Tab. 1

Dataset	Drivers	Passengers	Neutral	Disease	Total
CNO	3,163	-	3,163	-	6,326
Carcinoma	1,899	-	1,899	-	3,798
Haematopoietic	461	-	461	-	922
Lymphoid	441	-	441	-	882
Glioma	384	-	384	-	768
Melanoma	257	-	257	-	514
CND	3,163	-	1,581	1,582	6,326
Synthetic	3,163	3,163	-	-	6,326

Tab 1. Composition of datasets used in this project. They are composed by the same set of driver cancer variants and respectively only neutral polymorphisms (CNO), neutral and other disease-related variants (CND) and passenger cancer variants generated by CHASM algorithm (Synthetic).

The performances of the method on CNO and CND dataset and all the subsets relative to Carcinoma, Haematopoietic, Lymphoid, Glioma, Melanoma types of cancer are reported in Tab 2.

Dataset	Q2	P[D]	S[D]	P[N]	S[N]	C	AUC
CNO	0.93	0.93	0.93	0.93	0.93	0.86	0.98
Carcinoma	0.93	0.93	0.94	0.94	0.93	0.87	0.98
Haematopoietic	0.90	0.93	0.87	0.88	0.93	0.80	0.96
Lymphoid	0.93	0.93	0.92	0.92	0.93	0.85	0.98
Glioma	0.94	0.93	0.96	0.96	0.93	0.89	0.99
Melanoma	0.95	0.93	0.98	0.98	0.93	0.90	0.99
CND	0.90	0.87	0.93	0.92	0.86	0.79	0.95

Tab 2. Performances of the cancer-specific algorithm on the CNO and CND datasets, and on the Carcinoma, Haematopoietic, Lymphoid, Glioma, Melanoma cancer subsets. Q2 is the overall accuracy, C is the Matthews Correlation Coefficient, AUC is the area under the ROC curve P and S are Positive Predictive Values and Sensitivities for cancer-causing (D) and non cancer-causing (N) mSNVs.

In Fig.2 (panel A), we plot the ROC curves of SPF-Cancer on CNO and CND.

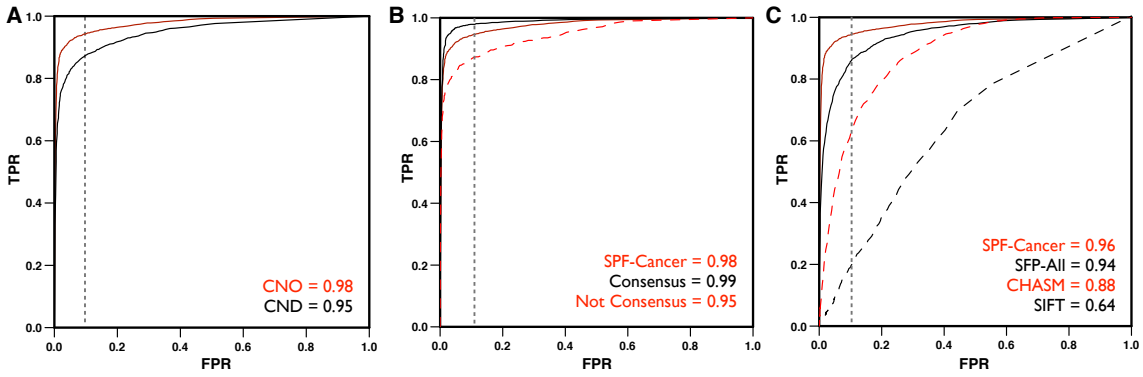

Fig 2. ROC curve of SPF-Cancer method on CNO and CND (panel A) on CNO dataset and Consensus and Not Consensus subsets (panel B). In panels C, ROC curves of SIFT, CHASM, SPF-All and SPF-Cancer methods on the Synthetic dataset.

We analyzed the results of the predictor considering the output of two different SVMs that take in input subset of SPF-Cancer input features. The first SVM (SVM-SEQPROF) takes in input only a 45-elements vector encoding for the mutation, the sequence environment and the profile. The second classifier (SVM-GOS) is based only on a 2-elements vector encoding for the protein function. Using this two binary classifiers we can filter the predictions of SPF-Cancer on the basis of their output. Thus, we calculate the performance of SPF-Cancer on two subset of the CNO dataset where the predictions of SVM-SEQPROF and SVM-GOS are in agreement or not. The results show that on the 62% of the CNO dataset on which the predictions of SVM-SEQPROF and SVM-GOS are in agreement the SPF-Cancer method reaches 96% of overall accuracy and 0.92 Matthews correlation coefficient. More details about the results are summarized in Tab. 3. The ROC curves of the SPF-Cancer method on the CNO dataset and its subsets are plotted in Fig. 2 (panel B).

Datasets	Q2	P[D]	S[D]	P[N]	S[N]	S[N]	AUC	PM
CNO	0.93	0.93	0.93	0.93	0.93	0.98	0.98	100
Consensus	0.96	0.96	0.95	0.96	0.97	0.92	0.99	62
NotConsensus	0.88	0.90	0.90	0.87	0.87	0.76	0.95	38

Tab 3. Performance of SPF-Cancer method on the CNO dataset and its subset on which SVM-SEQPROF and SVM-GOS agree (Consensus) and not (NotConsensus). Q2 is the overall accuracy, C is the Matthews Correlation Coefficient, AUC is the area under the ROC curve P and S are Positive Predictive Values and Sensitivities for cancer-causing (D) and non cancer-causing (N) mSNVs.

Finally, we compared the accuracy of SPF-Cancer method against the more general method (SPF-All) obtained using all disease-related mutations in training and against previously developed method like CHASM and SIFT. The results of

this test are reported in Tab 4. The ROC curves relative to the different methods are plotted in Fig. 2 (panel C).

Method	Q2	P[D]	S[D]	P[N]	S[N]	C	AUC	PM
SIFT	0.61	0.62	0.66	0.60	0.56	0.22	0.64	95
CHASM	0.80	0.85	0.73	0.76	0.87	0.60	0.88	100
SPF-All	0.88	0.88	0.87	0.87	0.88	0.75	0.94	100
SPF-Cancer	0.90	0.91	0.90	0.90	0.91	0.81	0.96	100

Tab 4.. Comparison between SPF-Cancer and other previously developed methods on the Synthetic dataset. Q2 is the overall accuracy, C is the Matthews Correlation Coefficient, AUC is the area under the ROC curve P and S are Positive Predictive Values and Sensitivities for cancer-causing (D) and non cancer-causing (N) mSNVs. PM is the percentage predicted variants for the Synthetic dataset.

The results show that SPF-Cancer performs better with respect to previously developed methods. It is interesting to notice that SPF-Cancer method, which uses a cancer-specific Gene Ontology (GO) score, is 2% more accurate than the SPF-All method, which uses all the functions to calculate the GO score.
The analysis and comparison of the GO slim-based scores used in SPF-Cancer and SPF-All methods show that cancer-specific GO term score better recognize particular protein functions related to cancer. In Fig. 3 we report the scatter plot of the logarithms of the cancer-specific versus the general GO-based scores.

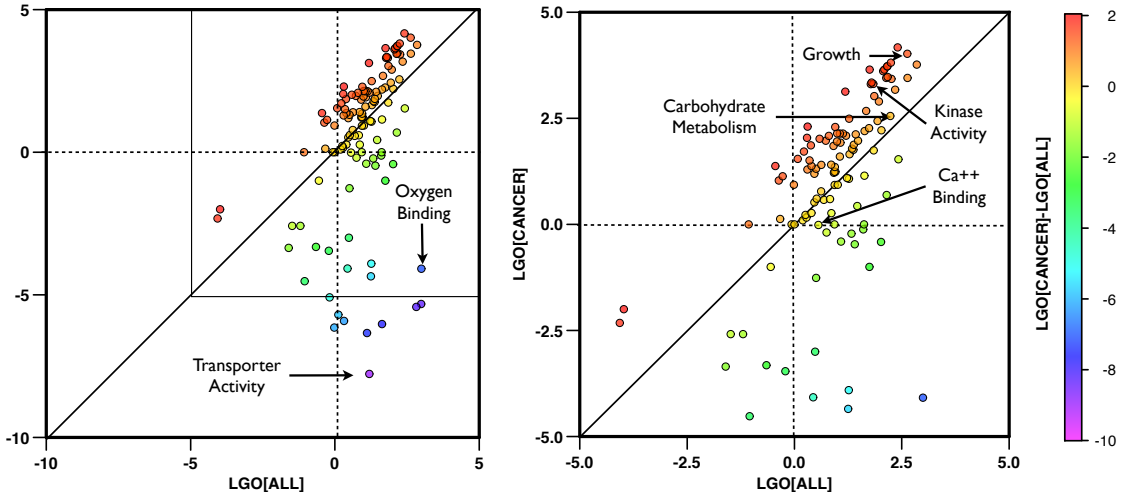

Fig 3. Scatter plot of the generic versus the cancer-specific LGO scores (LGO[All] and LGO[Cancer]) for each GO slim term (panel A). Color scale is related to the value of LGO[Cancer]- LGO[ALL]. In panel B, zoom of the plot in the region of LGO scores between -5 and 5.

The interesting GO functions are those corresponding to the points far from the diagonal. The points with negative generic LGOs and positive cancer-specific

LGOs are those with GO slim functions related to cancer. The points with cancer-specific LGOs close to zero and higher generic LGOs are those with GO slim functions generally associated to the all the pathologies in SwissVar dataset. For example, in our study we observed that Growth (GO:0040007) and Kinase Activity (GO:0016301) GO slim terms have stronger association to cancer showing respectively cancer-specific LGOs 4.02 and 3.30 and generic LGOs 2.63 and 1.78. Other interesting GO slim terms associated to all the diseases are the Transporter Activity (GO:0005215) and Oxygen Binding (GO:0019825) which have respectively cancer LGOs -7.77 and -4.09 and generic LGOs 1.20 and 2.99. There are also GO slim terms that have similar values for cancer and generic diseases LGO scores. Two examples are the Carbohydrate Metabolic Process (GO:0005975) that has similarly related cancer and all the diseases in our dataset resulting in LGO scores respectively 2.55 and 2.23, and the Calcium Ion Binding (GO:0005509) that is not related to cancer and slightly associated to all the diseases showing LGO scores -0.01 and 0.56 respectively.

1.2 World Wide Web server for the disease-related mutation prediction

During the last period of the returning phase, to accomplish the 6th task, EC implemented different web servers to make available to the scientific community the methods developed in this project. In detail,

WS-SNPs&GO server.

EC implemented an updated version of the SNPs&GO algorithm that predicts the effect of mSNVs using only sequence information. According to the findings of this research activity a new version of the SNPs&GO algorithm that takes in to account protein structure information (SNPs&GO^{3d}) has been made available on the web. SNPs&GO server (WS-SNPs&GO) with its implementations based on protein sequence and three-dimensional structure are reachable at http://snps.uib.es/snps-and-go.

The flow chart of the SNPs&GO server describing all the steps behind the prediction of the effect of mSNVs is represented in Fig 4.

Depending on the information available to the user, either SNPs&GO and/or SNPs&GO^{3d} can be activated. The server is endowed with two alternative input pages that are linked to the WS-SNPs&GO home page.

SNPs&GO input. The standard SNPs&GO server needs in input the protein sequence, its relative mutations and the functional annotations (see Fig 4 panel A). The input can be provided in three different ways: i) by pasting in the appropriate textbox area the protein sequence in FASTA or raw formats; ii) by uploading a file from the local machine; iii) by typing the SwissProt code. When the SwissProt code of the protein is provided, the server automatically assigns the associated GO terms in all the three subontologies (Biological Process, Cellular Component and Molecular Function) defined in the Gene Ontology. Alternatively, protein functional annotation can be provided using the appropriate

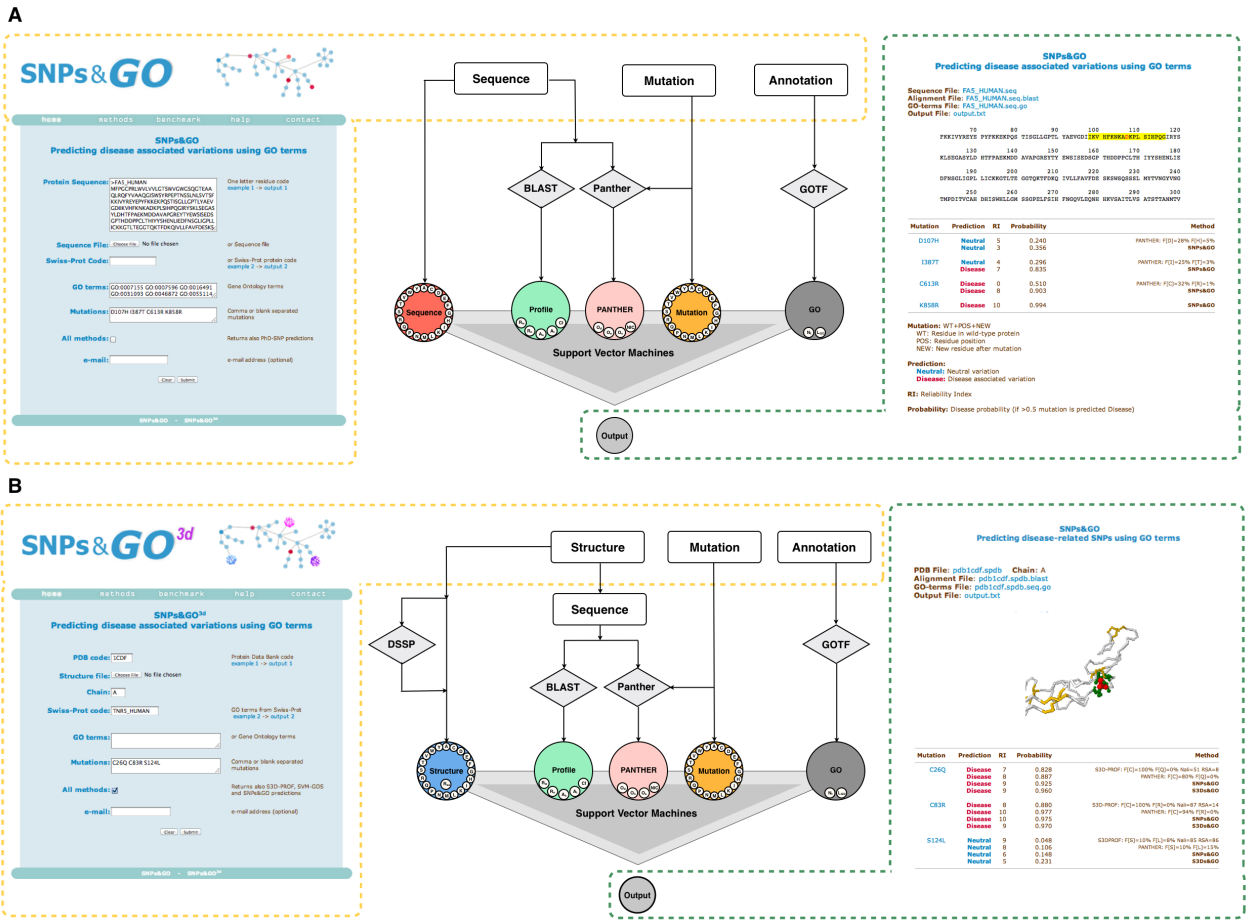

Fig 4. Schematic view of SNPs&GO (panel A) and SNPs&GO^{3d} (panel B). From the left to the right, the SNPs&GO and SNPs&GO^{3d} input web pages, the flow chart of the sequence and structure based methods and two examples of the returned outputs.

input box. In this case the server automatically runs the GO-TermFinder program (Boyle, et al., Bioinformatics 2004) for the retrieval of all the GO-term ancestors. When functional information is not provided the method assigns zeros to the two-elements vector encoding the protein function.

SNPs&GO^{3d} input. The SNPs&GO^{3d} interface (see Fig 4 panel B) is slightly different because in this case the server requires structural information. The input consists of: i) the PDB code (or a PDB file) of the mutated protein and the relative chain; ii) the list of mutations, iii) the protein GO terms. Also in this case, when the SwissProt code of the mutated protein is provided, the server automatically assigns all the annotation terms. More details about the input features are described in a previous work (Capriotti and Altman, BMC Bioinformatics 2011) performed during the outgoing part of the project.

WS-SNPS&GO output. The server has been designed to return the prediction output providing a link to a web page that is refreshed approximately each 20

seconds or by e-mail. The outputs of SNPs&GO and SNPs&GO^{3d} are similar. The html output page includes links to the sequence or structure given in input, to the results of the output of the BLAST search visualized with MView (Brown, et al., Bioinformatics 1998), to the file with all the GO terms associated to the mutated protein and the output in text format. In the second part of the output, the protein sequence is visualized and a table including all the mutations and their relative predictions is reported. In details, the table is composed of 5 columns. They are the mutated residue, the prediction (either Disease or Neutral), the reliability index (RI), the probability associated to the disease-related class and information about the prediction method. If the probability corresponding to disease-related is larger than 0.5 the variation is predicted as disease-related. In addition, a click on the variations in the output table, allows to highlight the mutated residue in the protein sequence visualized above. When available, the server also reports the output of the PANTHER algorithm (Thomas and Kejariwal, PNAS 2004), which is included in the input features of SNPs&GO. When the protein function is not available, the "All methods" option runs PhD-SNP (Capriotti et al., Bioinformatics 2006) and S3D-PROF (the 3D structure version of PHD-SNP). Both programs are based on sequence or structure profiles and the mutation environment. For SNPs&GO^{3d} the server returns outputs similar to those of SNPs&GO. The output includes also the Relative Solvent Accessible area (RSA) of the mutated residue calculated using the DSSP program (Kabsch and Sander, Biopolymers 1983). In the case of structural prediction the server exploits Jmol applet (http://jmol.sourceforge.net/) to visualize the protein structure and a click on the variation shows the mutated residue (in red) and its structural environment (in green). When the "All methods" option is activated the SNPs&GO^{3d} algorithm also returns the standard sequence-based SNPs&GO prediction.

Dr. Cancer server.
The promising results obtained in the analysis of cancer driver mutations have been used to implement a web server for predicting the cancer-causing mSNVs (Dr. Cancer). The Dr Cancer web server is available at
http://snps.uib.es/drcancer. The Dr Cancer web server uses similar architecture implemented for the sequence-based SNPs&GO algorithm but in this case the functional score is calculated using a specific set of proteins, which have at least one cancer-causing mSNVs. Given the smaller number of mutations included in the training set with respect to those used for the general method we used a reduced version of Gene Ontology, namely GO slim. The Dr. Cancer server is currently working, but some functionalities and help web pages are still under development.

Dr. Cancer input. The input required by Dr Cancer server is similar to that required by the sequence-based SNPs&GO servers. Dr. Cancer requires the

sequence of the protein, the mutation and functional information that can be provided by of the GO terms or automatically retrieved when the SwissProt ID of

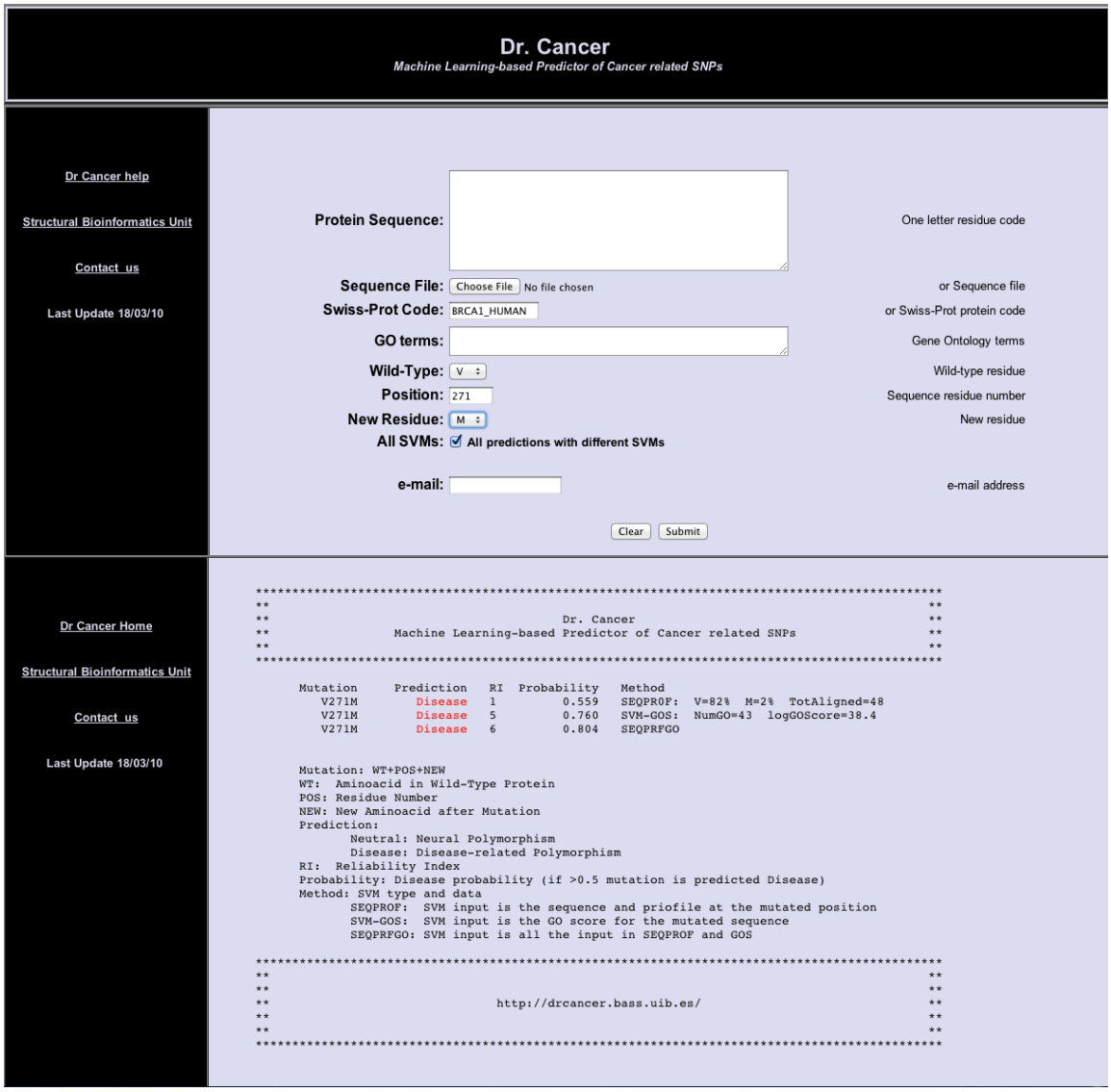

Fig 5. Input and output pages of the Dr. Cancer server. On the top of the figure is represented the input page of Dr. Cancer server with the field required for the predictions. On the bottom an example of output page with

the protein is given. On the top of Fig 5 we show the screen shot Dr Cancer's input page.

Dr. Cancer output. The developed tools behind the Dr Cancer web server automatically calculate information from the protein sequence. More in detail, the server executes a BLAST search to retrieve similar sequences and evaluates the degree of conservation of the wild-type residue in the mutated site. The algorithm also evaluates the occurrences of the 20 residues around the mutated position

thanking in to account a window of 19 residue centered on the mutated position. When the option of running all the SVM-based methods is checked the server returns predictions from the SVM-SEQPROF, SVM-GOS and SPF-Cancer algorithms. This option is important to select the subset of high reliable predictions on which all the 3 methods agree. In detail, for each mSNVs the server returns “Disease”, when the mutation is cancer-causing, or “Neutral” in all the other cases. The server provides other prediction-related measures, such as the reliability index and the probability to be a cancer-causing mutation, are reported to verify the reliability of the prediction. In the bottom part of Fig. 5 an example of Dr. Cancer’s output is shown.

2. Researcher training activities/transfer of knowledge activities/ integration activities.

In the period of the returning phase at University of Balearic Islands, EC was contracted researcher in the Department of Mathematics and Computer Science. EC attended the Bologna Winter School 2012, a 5-day course dedicate to the study of the proteins and their variants from the structural and functional point of view. He also had the opportunity to attend the course of Optimization held by Dr Jairo Rocha. There has been also the opportunity of collaboration with other members Computational Biology and Bioinformatics Research Group to perform a statistical analysis for the detection of high discriminative sequence and structure based feature included in our algorithms.

3. Highlight significant results

During the second phase, EC achieved many significant results related to the main aims of the Mut2Dis project. First of all, EC analyzed large dataset of cancer-causing mSNVs evaluating evolutionary and functional information to discriminate them from neutral polymorphisms and other disease-related mutations. The results have shown that residue conservation in the mutated site from the protein sequence profile is one the best discriminative features. This finding has been also verified comparing the subsets of cancer-causing mSNVs and polymorphisms. We have also shown that cancer-specific GO scores are more accurate that general GO-term ones in the identification of cancer-related protein, improving the detection of cancer-causing mSNVs. Finally, the new version SNPs&GO algorithm resulting from this research project has been scored between the best in its category either in testes performed by other groups (Thusberg et al. Human Mutation 2011) and in the blind set of mutations on CHK2 released by the Critical Assessment for Genome Interpretation (CAGI) organizers during the last two editions.

The great interest of the international scientific community on our methods is shown from the geographic (http://snps.uib.es/) and the numeric (http://snps.uib.es/awstats/awstats.pl) representations of the access to the http://snps.uib.es web server during the last few years.

4. Statement on the use of resources

For the development of this project during the returning phase the University of Balearic Island had the expenses reported in the following table.

Cost category	Expenses
Living and Mobility Allowance	50,615.00 €
Participation expenses	6,000.00 €
Travel allowance	250.00 €
Overheads (10%)	5,906.97 €
Management	2,204.70 €
TOTAL	**64,976.67 €**

The part of the funds dedicated to the project has been used to attend international conferences or to meet collaborators in US, Spain and Italy. Thus, EC disseminated the results obtained in the first period of the project, presenting his work at the Bologna Winter School 2012, the SNP-SIG 2012 and ISMB meeting 2012.
To have better opportunity to connect to the web and monitor our web servers everywhere during the trips a monthly USB contract for the connection was signed with SIMYO company in Spain. The total amount of expenses to support the Mut2Dis research project during the last year is 6,000.00 Euro (see the eligible costs table).