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	Outgoing Phase, Sep 2009 – Aug 2011 (24 months)

WORK PROGRESS AND ACHIEVEMENTS DURING THE PERIOD

1. Progress towards objectives and details for each task

In this section we summarized the objectives achieved for each one of the five aims described in our proposal during the outgoing phase at the Stanford University.

1.1 Study and characterization of the rate of evolution of Single Nucleotide Polymorphisms and their effect in human disease.

To achieve the first objective of our proposal Dr Capriotti did two different type of analysis. The evolution of the protein in the mutated position was studied considering both alignments of similar protein sequences or DNA alignment between homolog genes.
Thus, after collecting a dataset of annotated missense Single Nucleotide Variants (mSNVs), for each protein sequence we build a protein sequence profile aligning the mutated protein with all related proteins retrieved running BLAST on UniRef90 database with e-value lower than 10^{-9}. When the multiple sequence alignment is calculated, we compared the percentage of the wild-type and mutated residues in the mutated positions for the disease-related and neutral mSNVs. We found a significant difference between the two distributions (see Fig. 1 panel A). Similar difference has been found between the distributions of the Conservation Index (CI).

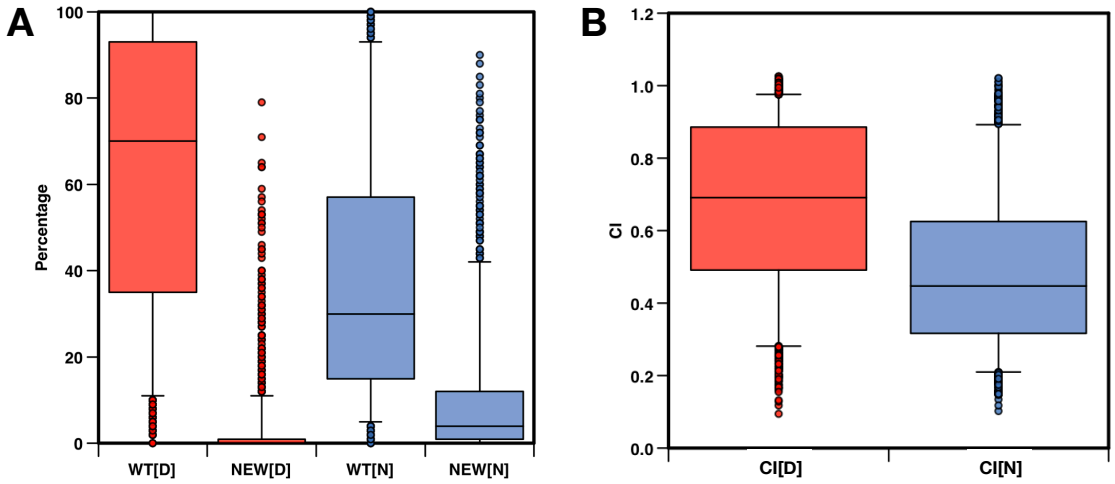


Fig. 1. In panel A, distribution of the percentages of wild-type (WT) and mutant (NEW) residues for disease-related (red) and neutral (blue) mSNVs. In panel B the distributions of the Conservation Index (CI) for the two classes of variants. D and N refer to disease-related and neutral variants.

In Tab. 1 we reported the median values of the distribution in Fig. 1.

	Disease	Neutral	KS p-value
WT	66	34	$>10^{-3}$
NEW	0	4	$>10^{-3}$
CI	0.66	0.47	$>10^{-3}$

Tab. 1. Median of the distributions of the percentages of wild-type (WT) and mutant (NEW) residues and conservation index (CI) for disease-related and neutral mSNVs.

Using different features, these data confirm the idea the wild-type residues are more conserved in disease-related sites than neutral and the mutated residues appear more frequently in the positions of the multiple sequence alignment corresponding to neutral mutation with respect to those disease-related. Similar results are obtained calculating the CI that is a position dependent measure, independent from the wild-type and mutant residues (see Fig 1 panel B).
Evolutionary information was also evaluated calculating the selective pressure acting at codon level for each mutated site. This task has been performed aligning the gene corresponding to the protein under mutation with ortholog genes in mammals. The alignments where performed using Muscle tool and selective pressure acting at codon level were calculated using the likelihood model M2 implemented in the PAML package. This procedure is more time consuming than standard BLAST alignment and in many cases, when a gaps are present in the mutated position of the alignment or less than 4 orthologs are available the selective pressure can not be calculated. Thus, we calculated the omega values, which is the ratio between the rates of non-synonymous versus

synonymous variations (ω=dN/dS), only for the subset of variants for which all the evolutionary information were available. Starting from a dataset of more than 55,000 annotated mutations we obtain omega values for 25,237 mutated sites corresponding to 6,224 proteins. In Fig. 2 we report the boxplot of the distributions of the ω for disease-related and neutral variants

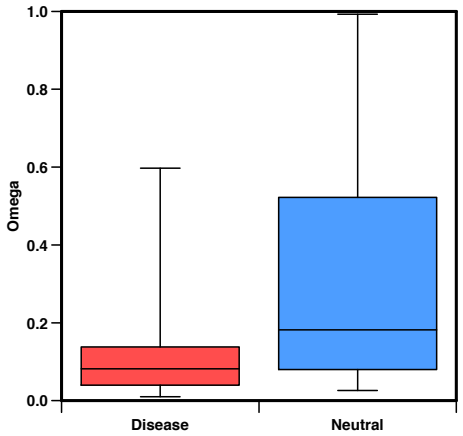


Fig. 2. Distributions of the ω values of the mutated positions for disease-related (red) and neutral (blue) variants.

The results show that the median of the distribution of the ω values are 0.08 and 0.18 for disease-related and neutral mutations respectively.

1.2 Study and characterization of the structural determinants of human disease.

After the characterization of mSNVs in terms conservation and evolution across species, we analyzed structural features of mutated residues. In this direction the first goal has been the development of an automatic method to map the mSNVm for protein sequence to structure. The mapping tool performs a BLAST alignment of the protein sequences under mutation and the reference sequences of the protein collected in the Protein Data Bank (PDB). To select only mSNVs with perfect overlap, we only consider alignments with 100% sequence identity, no gaps and alignment length more than 39 residues. When a variant maps with more than 1 structure, the one with best resolution is selected. This filtering procedure reduces sensibly the number of variants in the original dataset but it provides a clean and large subset to perform protein structural analysis. Thus, we mapped in total 4,986 mSNVs of which 3,342 are disease-related and 1,644 neutral. The whole set of variants are associated to 784 PDB chains. The protein structures and the mutated sites in the selected dataset have been analyzed to find new discriminative features. First we calculated the distributions of the relative solvent accessible area (RASA) for disease-related

and neutral mSNVs and the occurrences in each secondary structural class (see Fig. 3).

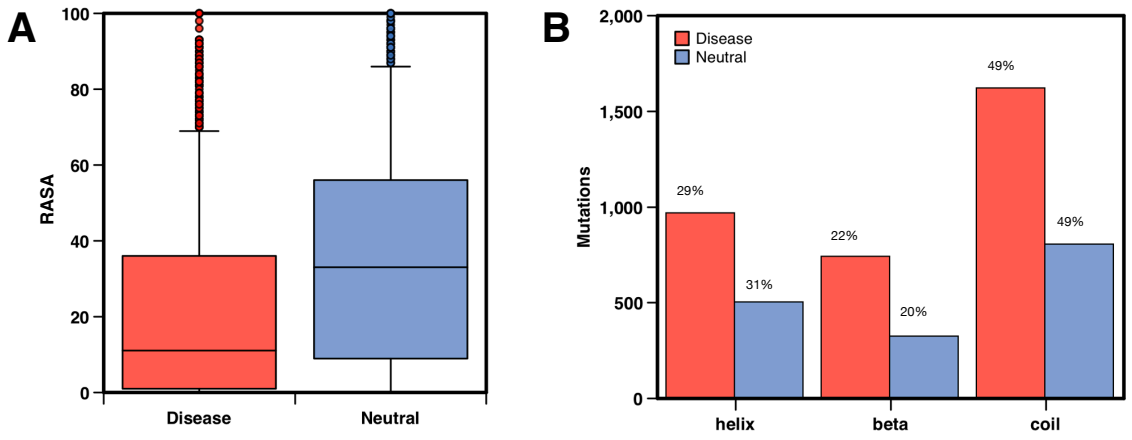


Fig. 3. In panel A, distribution of the relative solvent accessible area (RASA) of the wild-type residue for disease-related (red) and neutral (blue) mSNVs. In panel B, the number of wild-type residue in each one of the secondary structure classes (helix, beta and coil).

The results show a significant difference between the distributions of RASA for disease-related and neutral variants. More in details the disease-related mutations are more likely to happen in the core of the proteins and neutral ones on the surface. According to this the median values of RASA are about 20 and 36 for disease-related and neutral variants respectively. On this subset of mutations no significant difference between the occurrences of disease-related and neutral classes in the different secondary structure categories has been detected.

1.3 Development of new general machine learning methods for disease prediction.

The data extracted according to the procedure described above have been used to train and testing machine learning approach to classify mutation in disease-related and neutral polymorphisms. In particular EC first developed a sequence-based method. This algorithm is a Support Vector Machine trained on data extracted from protein sequence, profile, the output of PANTHER and a functional score derived using Gene Ontology terms of the protein under mutations and all their parents. The sequence-based SVM takes in input a 51 elements vector including:

- 20 elements vector encoding for the mutation;
- 20 elements vector for the sequence environment;
- 5 elements vector for the sequence profile;
- 4 elements vector from PANTHER output;
- 2 elements vector for the functional score.

A second SVM base method has been developed including structural information. The structure-based algorithm takes in to account the Relative Accessible Solvent Area of the mutated residue and considers the structural environment of the mutated position. More in details the sequence-base method calculates the frequencies of the amino acids in a window of 9 residues each side around the mutated residue while the structure-based method calculates the frequencies of the amino acids falling in a radius shell of 6 Å from the Cα of the mutated residue. Differences between the two methods are represented in Fig. 4.

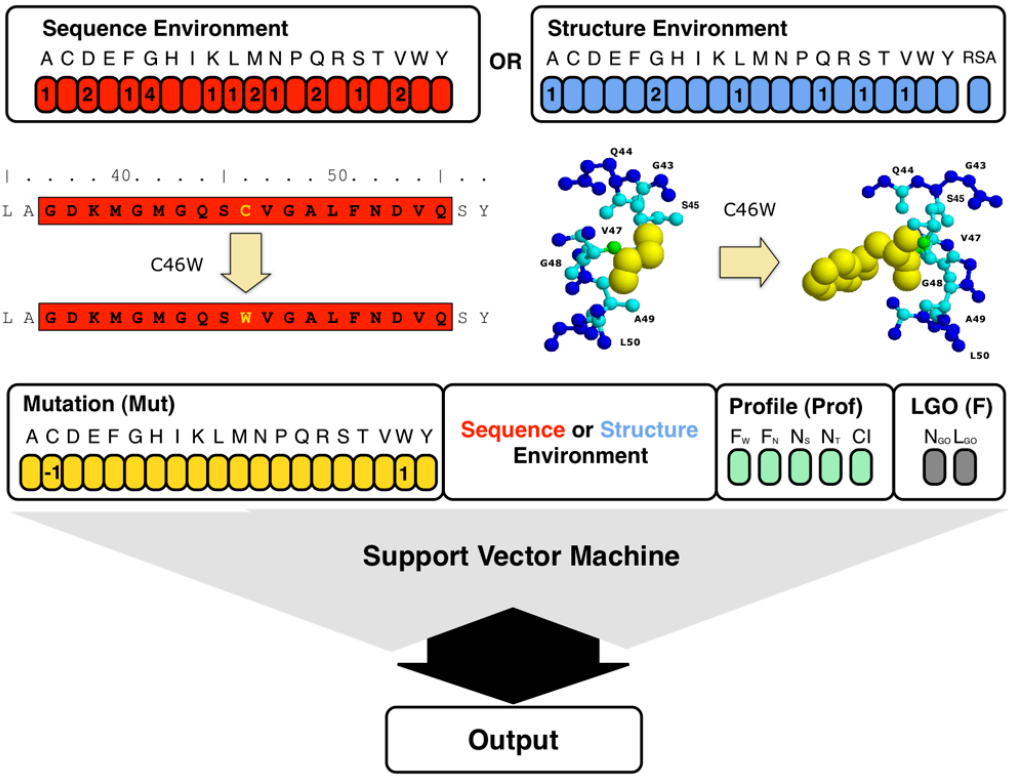


Fig. 4. Flow chart of our SVM-based methods. The structure-based method (SVM-3D) takes in input mutation (yellow) structure environment (in blue), sequence profile (green), PANTHER output (pink) and function (gray) information. In the sequence-based method (SVM-SEQ) the 21 elements vector encoding for the structural environment is replaced by the 20 elements vector encoding for the sequence environment. The structure environment is the residue composition in a 6 Å radius shell around the C-α of the mutated residue. The sequence environment is the amino acid composition window of 19 residues centred on the mutated residue.

Thus the sequence-base algorithm has been used as baseline to quantify the improvement of the prediction resulting from the use of structural information.
After a 20-fold cross-validation procedure we compared the accuracy of the structure-based method (SVM-3D) against the sequence base one (SVM-SEQ). The results are reported in table 2.

	Q2	P[D]	S[D]	P[N]	S[D]	C	AUC
SVM-SEQ	0.82	0.81	0.83	0.82	0.81	0.64	0.89
SVM-3D	0.85	0.84	0.87	0.86	0.83	0.70	0.92

Tab. 2. Performances of the sequence (SVM-SEQ) and structure (SVM-3D) based methods. Q2, P and S are overall accuracy, Precision and Sensitivity respectively. C and AUC are the Matthews Correlation coefficient and the Area Under the ROC Curve. D and N stand for disease-related and neutral variants.

In figure 4 I plot the AUCs of each method and the accuracy of the SVM-3D method as a function of the reliability index (RI)

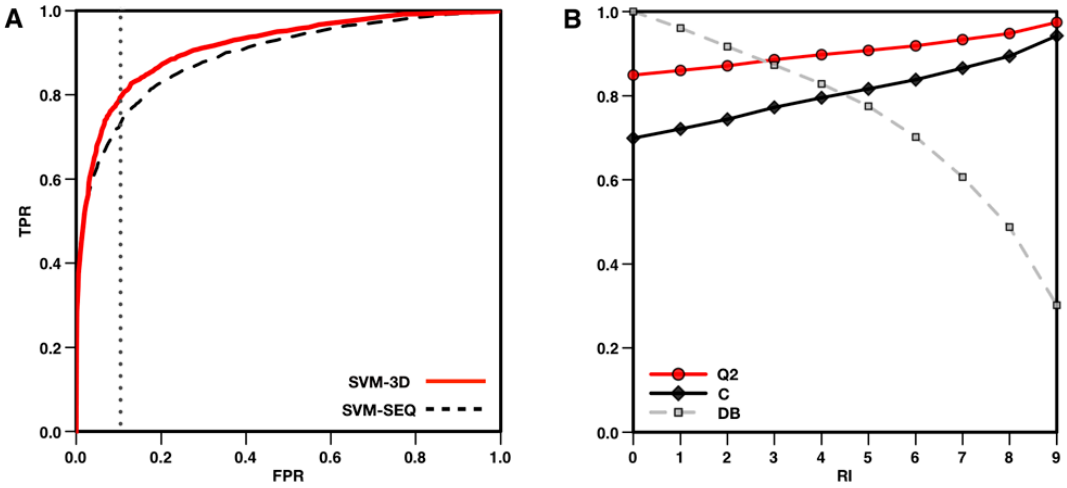


Fig. 4. Performance of the structural-based method. In panel (A), ROC curves of the sequence (SVM-SEQ) and structure-based methods (SVM-3D). The plot shows the improvement of 3% in AUC and 7% in TPR when sequence and structure base methods are compared. In panel B, accuracy and correlation coefficient of SVM-3D as function of the Reliability Index (RI). If predictions with RI>5 are selected the SVM-3D method results in 91% overall accuracy 0.82 correlation coefficient over 78% of the dataset. Accuracy measures (Q2, C, TPR and FPR) are defined in Methods section. DB is the fraction of the whole dataset of mutations.

The results have shown that the structure-based method results in 3% improvement in overall accuracy and 0.06 higher correlation. Comparing the ROC curves (Fig. 4 A), SVM-3D gives 0.03 better Area Under the Curve (AUC) with respect to SVM-SEQ. If 10% of wrong predictions are accepted, SVM-3D has 7% more true positives. The output returned by the SVM has been used to calculate the Reliability Index (RI) in order to filter predictions. If predictions with RI>5 are selected, the SVM-3D method achieves 91% overall accuracy and 0.82 correlation coefficient on 78% of the whole dataset (see Fig. 4 B).

1.4 Development of disease-specific predictors

For this particular task EC is focusing on the detection of mutations involved in the insurgence of different types of cancers. Thus a set of manually curated driver variants associated to cancer has been selected. For this particular task we have planned to develop a machine learning specific for the detection of cancer-causing mutations. The largest part of this task will be performed in the retuning phase.

1.5 Development of a World Wide Web server for predicting the likelihood of a SNP variant to be associated with human disease.

Will be developed during the returning phase (see research plan).

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

In the period of the outgoing phase at Stanford University, EC was appointed as postdoc in the Department of Bioengineering. He had the opportunity to attend different courses. In particular during the first year EC attended the courses BIOMEDIN212 (Fall 2009) learning how to write and present a NIH research grant. In winter 2010 EC attended the BIOMEDIN211 courses learning how to design appropriate systems for medical applications. In summer 2010, EC had the opportunity to watch the recorded classes of the BIOMEDIN214 course available only online for this trimester. In this course EC learned advanced techniques for protein sequence and structure analysis and python programming. During the second year EC listened the recorded classes of the BIOCHEM228 courses available only online during Fall 2010. With this course EC had the learned about advanced methods for sequence analysis in particular for the detection of protein motifs. During the winter 2011 EC attended the GENETICS211 course learning about the biochemistry and theory behind standard techniques used in Genomics and Proteomics. Unfortunately the BIOE331 course described in the proposal as interesting for EC's training was not attended because not given during the first part of 2011. For the courses physically attended al letter of the teachers is included in this report. In 2010, EC attended the 4th annual Comprehensive Cancer Research Training Program where he learned about the most advanced methods for cancer diagnostic and treatments. During the fist year at Stanford, EC also attended specific courses for postdocs to learn about the negotiation of a group leader position and how lead a research group and other course about grant opportunity, application and management. At Stanford, EC worked in the Helix group directed by Dr. Russ Altman thus he had the opportunity to collaborate with undergraduate and graduate students in Bioengineering and Biomedical Informatics. In addition he

cooperated for specific projects with the researcher at the PharmGKB consortium that is interested to analyze the effect of genetic variants on drug response. Thus EC learned ho to use the tools implemented for the retrieval information on PharmGKB database that links genes to drug and diseases.

3. Highlight significant results

During the first phase, EC achieved many significant results related to the main objectives of the Mut2Dis project. First of all, EC analyzed large dataset of annotated mutations evaluating evolutionary and structural information. Using higher number of variants, he confirmed hypothesis published in a previous paper that evolutionary information (Capriotti et al., Human Mutation, 2008, PMID: 17935148) and functional information (Calabrese et al., Human Mutation, 2009, PMID: 19514061) are important for the detection of deleterious variants. The analysis of protein structure information has been used to select important features to improve the prediction of disease-related mutations. This improvement has been quantified comparing the results reached using binary classifiers based on protein sequence and structural information. More practical results is the development of a new machine learning base approach based on protein structure information to predict the effect of missense single nucleotide variants. All the details about this study have been recently published (Capriotti and Altman, BMC Bioinformatics, 2011, PMID: 21992054).
In addition the experience accumulated in this topics during the last to years have allowed to EC to collaborate in the publication of a Review about the bioinformatics challenges in personalized medicine (Fernald et al. Bioinformatics 2011, PMID: 21596790). Finally, the methods developed by EC have been used to predict the effect of a blind set of mutations released for the Critical Assessment of Genome Interpretation experiments in the 2010. After evaluation, EC's tools were scored between the best in the prediction of deleterious mutations thus EC has been selected for a short presentation during the final workshop at the UC Berkeley. During his period at Stanford, EC had also the opportunity to meet in person internationally recognized researcher interested in the development of new methods for genome interpretation allowing him to establish new collaboration with Sean Mooney at Buck Institute, Novato (CA), Inna Dubchak, Lawrence Berkeley National Laboratories, Berkeley (CA), Yana Bromberg, Rutger University New Brunswick (NJ).

4. Statement on the use of resources

For the development of this project during the last two years the University of Balearic Island had the expenses reported in the following table.

Cost category	Expenses
Living and Mobility Allowance	118.590,00 €
Participation expenses	13.801,12 €
Travel allowance	4.163,49 €
Overheads (10%)	13.655,46 €
Management	6.265,07 €
TOTAL	**156.475,14 €**

In the Form C attached to this report we included only the amount provided by the European Community (Total 152,627.84 €).
The participation expenses funds of this project have been used mainly to buy a laptop computers used to implement and run the programs used for our analysis and for participation to meeting and conferences.
In particular a laptop computer Dell has been bought for a total amount of 1,400.52 Euro and it is personally used by EC during the development of the project. The part of the funds dedicated to the project has been used to attend international and national 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 ISMB2010 and 2011, the ECCB2010, the PSB2011, and so on. Smaller part of the budget has been used to pay short trips to visit collaborators in Berkeley and Novato.
To have better opportunity to be connected on the web everywhere during the trips a monthly USB contract for the connection was signed with Verizon Wireless company in US. The total amount of expenses to support the Mut2Dis research project during the first 2 year is 12,000.00 Euro (see the eligible costs table).

Russ Biagio Altman, MD, PhD
Professor & Guidant Chair of Bioengineering
Director, Biomedical Informatics Training Program
Professor, Genetics & Medicine
Professor (by courtesy), Computer Science
Clark Center S172, 318 Campus Drive
Stanford University, Stanford, CA 94305-5444
Tel: 650-725-3394 Fax: 650-723-8544
russ.altman@stanford.edu

January 26, 2010

Subject: BIOMEDIN 212 course attendance

To whom it may concern,

I am writing this letter to certify that **Dr. Emidio Capriotti** attended and completed the BIOMEDIN 212 course at the Stanford University during the fall quarter in the 2009. This course is entitled "Introduction to Biomedical Informatics Research Methodology."

Sincerely,

Russ B. Altman

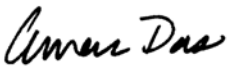
STANFORD CENTER FOR
BIOMEDICAL INFORMATICS RESEARCH

May 20, 2010

To Emidio Capriotti:

Thank you for your participation in my course, BIOMEDIN 211: “Effective Design in Clinical Informatics.” This letter is to confirm that you attended the course lectures as an auditor and that you successfully completed the homework assignments.

Sincerely,



Amar K.Das, M.D., Ph.D.
Assistant Professor
Department of Medicine (Biomedical Informatics)
Stanford University

Stanford Center for Biomedical Informatics Research
Medical School Office Building, Room X-215; 251 Campus Drive, Stanford, CA 94305-5479
Phone: 650-723-0679 · Fax: 650-725-7944 · Web: http://bmir.stanford.edu/

department of GENETICS

STANFORD UNIVERSITY
SCHOOL OF MEDICINE

Gavin Sherlock, Ph.D.
Associate Professor
sherlock@genome.stanford.edu

March 23rd 2011

To Whom It May Concern:

I am writing to confirm that Emidio Capriotti audited GENE211, which is a course taught by Prof. Mike Cherry and myself on Genomics. The syllabus for the course is available at:

http://gene211.stanford.edu/schedule.html

Emidio attended lectures and participated in the discussion sections. Please don't hesitate to ask if you need more information.

Sincerely,

Gavin Sherlock

DEPARTMENT OF GENETICS
GRANT BUILDING, RM S201A, 300 PASTEUR DRIVE, STANFORD, CA 94305-5120
TEL: 650.498.6012 FAX 650.724.3701