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)

RESEARCH SUMMARY

1. Summary of the project objectives

In this report we summarize the research activity performed by Dr. Emidio Capriotti during the outgoing phase of the Marie-Curie IOF at the Department of Bioengineering, Stanford University under the supervision of Dr. Russ B. Altman.
The main aims of our proposal are the following:

i. Study and characterization of the rate of evolution of Single Nucleotide Polymorphisms and their effect in human disease.
ii. Study and characterization of the structural determinants of human disease.
iii. Development of new general machine learning methods for disease prediction.
iv. Development of disease-specific predictors
v. Development of a World Wide Web server for predicting the likelihood of a SNP variant to be associated with human disease.

These 5 aims correspond to 6 different tasks that have to be accomplished in 36 months. In the proposal's timeline, we planned to perform about 4 over the 6 tasks during the outgoing phase (24 months).
According to this, we mainly achieved the first 3 objectives and part of the 4th and 5th. The remaining parts of specific aims 4 and 5 will be performed during the returning phase at the University of Balearic Islands (Spain).

2. Description of the work performed since the beginning of the project

During the first months of the project Dr. Emidio Capriotti selected a set of annotated missense Single Nucleotide Variants (mSNVs) from the database SwissVar. The dataset used in this work has been downloaded at the end of October 2009. After a filtering procedure to remove unclassified mSNVs, we collected a dataset composed from 55,131 variants from 11,657 human proteins, 20,879 of which are classified as disease-related and 34,258 as neutral polymorphisms.
The selection of the subset of mSNVs for which the three-dimensionl structure of the proteins is known, EC implemented programs able to automatically compare the sequences of the mutated proteins with the sequences of the protein collected in the Protein Data Bank (PDB). We apply a very strict selection procedure selecting only proteins with complete sequence overlap and with length higher than 39 residues. Using this criteria, we collected a subset of 4,986

mSNVs from 784 PDB chains. This subset is composed by 3,342 disease-related and 1,644 neutral mSNVs. To accomplish the second objective of our proposal, EC analyzed protein structure around the mutated site to find the best features to discriminate between disease-related and neutral variants.
When the dataset of annotated mSNVs was available, the analysis of residue conservation in the mutated position has been performed aligning the protein under study with a set of proteins with high sequence similarity. These proteins have been selected running the BLAST algorithm on the UniRef90 database and including only hits with e-value lower than 10^{-9}. In a second phase EC performed a preliminary evolutionary analysis calculating the selective pressure acting at codon level using alignments between the human DNA sequences and their homolog in mammalian species.
In the next step Dr. Capriotti built a machine-learning base approaches to predict the impact of mSNVs evaluating the discriminative power of different features. In these algorithms we included features from sequence analysis such as evolutionary and functional information and protein structure information. In last period a disease-specific method have been developed to predict the cancer causing mSNVs. This algorithm has been built analyzing a manually curated set of cancer driver variants recently used to train a method for the discrimination between driver and passenger cancer mSNVs.

3. Description of the achieved results

With the research activity performed during the outgoing phase, EC reached the largest part of the objective described in our proposal. In particular, we defined a set of discriminative features derived from protein sequence profile and protein structure. We found that the distribution of the frequencies of the wild-type and mutant residues in the mutated sites for disease-related and neutral mSNV are significantly different. Analyzing the protein structure, we have shown strong differences between the distributions of the relative solvent accessible area for the disease-related and neutral variants.
These findings allowed to develop a new machine learning based method for the prediction of deleterious variant taking in input information from protein sequence profile, protein function and protein structure. The improvement of the prediction accuracy resulting from the use of structure information has been quantitatively estimated comparing the structure-based method with similar sequence-based tool. The results shown that the structure–based method is 3% more accurate than sequence-base method increasing the overall accuracy from 82% to 85%. In addition the structure-based information are able to provide more information about the biochemical mechanism of the disease.

4. Expected final results and their potential impact

After the returning phase we expect to have developed a user-friendly web server interface for the prediction of the effect of mSNVs. Currently, EC is implementing these web tools including both protein sequence and structure information. At the

same time Dr. Capriotti is developing the first disease-specific predictor for the detection of cancer-causing variants. Similar approach will be use to implement methods for other classes of disease. A the moment the preliminary version of this algorithms have been used participate in the CAGI experiments for genomic interpretation. We believe that the use of structural information in the prediction of deleterious variants will be important for the understanding of the disease mechanism. In addition the developed method will have an impact in personal genomics allowing to make new hypothesis about the insurgence of genetic diseases. As natural consequence of this work we are planning to study the relationship between genetic variants and drug response. The application of newly developed tools in clinical settings will be important for the establishment of personalized medicine.