Human-comprehensible rule generator for identifying deleterious amino acid variants

Tien-Dao Luu, Ngoc-Hoan Nguyen, Anne Friedrich, Jean Muller, Luc Moulinier and Olivier Poch

Laboratoire de BioInformatique et Génomique Intégratives de l'IGBMC (CNRS – UMR 7104), 1 rue Laurent Fries, Illkirch 67404, Strasbourg France

{luudao,nguyen@igbmc.fr}

Objectives

•Main goal: discover relationships between genotypic and phenotypic variations.

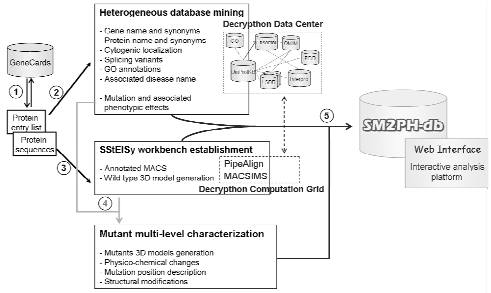
•Using Inductive Logic Programming (ILP) [1] to characterize and predict the impact of mutations on protein function in the context of the SM2PH-db ("from Structural Mutation to Pathology Phenotypes in Human database") [2]

Website: http://decrypthon.igbmc.fr/sm2ph/

(Journal of Medical Genetics)

SM2PH-db

SM2PH-db automated workflow for data generation and integration



Methods

Step 1 : Requirements definition and example construction

We confine this study to the task of discriminating **deleterious mutations** (positive examples) from **neutral mutations** (negative examples).

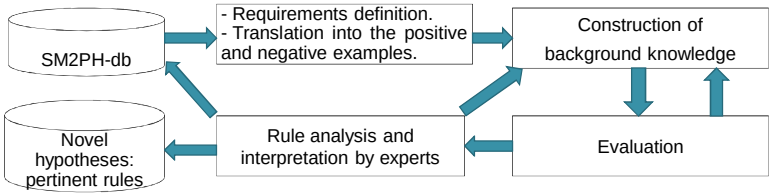
Dataset	Data ratio	Deleterious mutations	Neutral mutations	Total
DS1	4.88 : 1	6480 (83%)	1637 (17%)	8117
DS2	2:1	3274 (67%)	1637 (33%)	4911
DS3	1:1	1637 (50%)	1637 (50%)	3274

Step 2 : Background knowledge construction

Table : Predicates used as background knowledge

Levels of information	Predicates
Physico-chemical changes induced by the substitution	Modification size / charge / polarity / hydrophobicity / Gly or Pro / score
Functional and structural features	Conservation in the alignment
	Number of known mutations at this position
	Wild type/Mutant residue representation in the alignment
	In a secondary structure element?
Structural modifications induced by the substitution, based on the mutant models	Additional contact / Lost contact / Identical contact
	Additional contact n+1 / Lost contact n+1 / Identical contact n+1
	Wild type/mutant relative accessibility
	Grouping the mutations in the 3D cluster of 10A°
	DDG reliability
	I-mutant DDG variation

KDD process



Step 3 : Performance measurement and evaluation

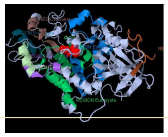
10 fold cross validation

	ILP			SIFT [3]			PolyPhen [4]		
Dataset	Se	Sp	Gm	Se	Sp	Gm	Se	Sp	Gm
DS1	85.97	52.54	67.14	71.72	67.31	69.48	78.05	64.27	70.82
DS2	77.74	66.76	71.96	71.90	67.31	69.57	77.88	64.27	70.75
DS3	70.39	75.23	72.96	71.59	67.31	69.42	80.57	64.27	71.96
Average	78.03	64.84	70.59	71.74	67.31	69.49	78.83	64.27	71.18

$$Se = \text{Sensitivity} = \frac{TP}{TP+FN} \qquad Sp = \text{Specificity} = \frac{TN}{TN+TP} \qquad Gm = Gmean = \sqrt{Se * Sp} \quad [5]$$

Step 4 : Rule analysis and interpretation by experts

```
deleterious(A) :-
    conservation_class(A, global_conservation_Rank1),
    conservation_wt(A, B) and B>=96.55,
    stability(A, decrease).
```



This rule states that a mutation A is deleterious if:

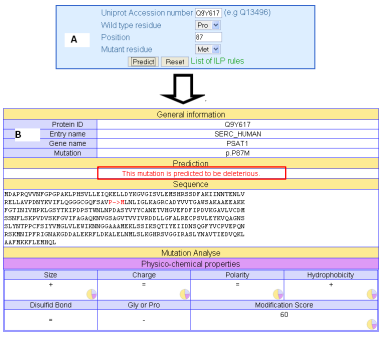
•The mutated residue is ranked in the "global conservation rank1" and

•Conservation in the wild type residue representation in alignment is greater than 96.55% and

•Stability of protein after point mutation is decreased.

Prediction service implementation

http://decrypthon.igbmc.fr/sm2ph/cgi-bin/prediction



Screenshots of prediction pages. (A) Input form. The mutation predicted in this example is the P87M of Phosphoserine Aminotransferase protein. (B) Output page which provides prediction result as well as multi-level (physico-chemical, functional, structural and evolutionary) characterizations of mutation.

Conclusions & Perspectives

✓This study presents a novel application of ILP in the bioinformatics domain, namely, characterizing and predicting the effect of mutation on protein function.

✓We have harvested a mutation knowledge base: set of rules and the important predictors for identifying deleterious mutations.

✓In the future, we plan to enrich the background knowledge by including: more detailed genotypic and phenotypic information, interactomic data such as functional and physical interactions mined from the STRING, KEGG, MPI…

✓We will also ameliorate ILP models to help doctors / biologists to understand the consequences of mutations at several levels:

- ▪Various degrees of severity: very severe, severe, intermediate, moderate, neutral, ...
- ▪Complex phenotypic descriptions: loss of walking ability, mental retardation, …

References

[1] Muggleton, S. (1991) Inductive logic programming, New Generation Computing, 8, 295-318.
[2] Friedrich, A., et al. (2009) SM2PH-db: an interactive system for the integrated analysis of phenotypic consequences of missense mutations in proteins involved in human genetic diseases, Hum Mutat.
[3] Ng, P.C. and Henikoff, S. (2003) SIFT: Predicting amino acid changes that affect protein function, Nucleic Acids Res, 31, 3812-3814.
[4] Ramensky, V., et al. (2002) Human non-synonymous SNPs: server and survey, Nucleic Acids Res, 30, 3894-3900.
[5] Kubat, M., et al. (1998) Machine Learning for the Detection of Oil Spills in Satellite Radar Images, Mach. Learn., 30, 195-215.

Acknowledgements: This work was funded by the Vietnam Ministry of Education and Training, the Institute National de la Santé et de la Recherche Médicale, the Centre National de la Recherche Scientifique (CNRS), the Université de Strasbourg, and the Décrypthon program initiated by the Association Française contre les Myopathies and IBM.