

Optimisation of Parameters for the Assessment of Unclassified Disease Gene Sequence Variants

Jennifer D Warrender
Daniel Swan
Ciaron McAnulty

Introduction

- What are the best performing PON tools
- For those PON tools that require a MSA
 - What is the best tool for creating a MSA
- For those PON tools that generate a confidence score
 - What is the best threshold to define a pathogenic mutation or benign polymorphism
- At the time of starting the project there was no large scale comparison of PON tools and there remains no published data that take into account 2 different MSAs.



Approach

- Investigation into which of the seven MSA tools gave the best alignment by using known MSA benchmarks
- Obtaining control data – known variants and sequences from Uniprot database
- Investigation of six PON tools and the best thresholds for AlignGVGD and PolyPhen
- Investigation into the optimal MSA(s) for each PON tool and the idea of a generic optimal MSA



PON Tools

- Used to:
 - Predict the potential harmful effects of nonsynonymous missense mutations in humans
 - Classify unknown missense variants
- Two types of PON tools:
 - Sequence based tools – evolutionary conservation of sequences which is dependent on the quality of the alignment
 - Structure based tools



PON Tools

- Allows Custom MSA Input
 - AlignGVGD
 - MAPP
 - PMut
 - SIFT

- Automatic MSA Generation
 - PolyPhen
 - SNPs&GO



Special Considerations

- PON tool input
- Consider the sequence source
- Consider the variation source
- AlignGVGD, PMut, SIFT and SNPs&GO will in general always produce a prediction.
- Possible for MAPP and SIFT to not produce a prediction for certain positions in the human sequence



MSA Tool Ranking

BAlIbASE

MSA	TC
T-Coffee	0.617
MUSCLE	0.572
ClustalW	0.554
MAFFT	0.551
Dialign	0.488
SAM	0.236
HMMER	0.157

PREFAB

MSA	TC
T-Coffee	0.709
MUSCLE	0.679
ClustalW	0.679
MAFFT	0.617
Dialign	0.599
SAM	0.475
HMMER	0.441



Alignments

- Every possible MSA combination of size 3-10
- Creates subsets of lengths 2 to 9 species (excluding *H. sapiens*)
- This resulted in a range of MSAs for each data set (between 26 and 502)
- Used to investigate if the MSA affects the PON prediction



PON Benchmark

- Use of Matthews Correlation Coefficient (MCC)
- Incorporates the false and true predictions
- Known for balancing the number of pathogenic and neutral predictions
- MCC scores are between -1 and +1
 - -1; 0% of predictions are correct
 - +1; 100% of predictions are correct



PON Thresholds

- AlignGVGD predictions were classified as C0; C15; C25; C35; C45; C55; C65 where C0 is benign and C65 pathogenic
- PolyPhen predictions were classified as Predicted or Possibly pathogenic and Benign
- Investigated if changing the threshold affects the PON prediction



PON Threshold Results

AlignGVGD

	Threshold	MCC
ABCC8	0-123456	0.700
CFTR	0-123456	0.173
F8	0-123456	0.350
FBN1	0123-456	0.292
GCK	0-123456	0.810
KCNJ11	012-3456	0.492
VWF	0-123456	0.530

PolyPhen

	Threshold	MCC
ABCC8	PP-B	0.565
CFTR	PP-B	0.080
F8	P-PB	0.212
FBN1	PP-B	0.049
GCK	PP-B	0.690
KCNJ11	P-PB	0.573
VWF	PP-B	0.501



PON Tool Results

	ABCC8		KCNJ11	
SIFT	1.000	100.00	0.580	79.00
AGVGD	0.700	85.00	0.492	74.60
MAPP	0.700	85.00	0.304	65.20
PolyPhen	0.565	78.25	0.573	78.65
PMut	0.484	74.20	0.400	70.00
SNPs&GO	0.286	64.30	0.573	78.65

	CFTR		GCK	
AGVGD	0.173	58.65	0.810	90.50
SIFT	0.173	58.65	0.810	90.50
PMut	0.163	58.15	0.554	77.70
MAPP	0.149	57.45	0.487	74.35
PolyPhen	0.080	54.00	0.690	84.50
SNPs&GO	0.000	50.00	0.000	50.00



PON Tool Results

	F8		FBN1	
MAPP	0.389	69.45	0.301	65.05
SIFT	0.351	67.55	0.211	60.55
AGVGD	0.350	67.50	0.292	64.60
PolyPhen	0.212	60.60	0.049	52.45
PMut	0.167	58.35	0.271	63.55
SNPs&GO	0.092	54.60	0.123	56.15

	VWF	
PMut	0.579	78.95
AGVGD	0.530	76.50
SIFT	0.529	76.45
MAPP	0.517	75.85
PolyPhen	0.501	75.06
SNPs&GO	0.330	66.50



PON Tool Ranking

	AGVGD	MAPP	PMut	PolyPhen	SIFT	SNPS&GO
ABCC8	2	2	5	4	1	6
CFTR	1	4	3	5	1	6
F8	3	1	5	4	2	6
FBN1	2	1	3	6	4	5
GCK	1	5	4	3	1	6
KCNJ11	4	6	5	2	1	2
VWF	2	4	1	5	3	6



Summary

- T-Coffee appears to provide the best MSAs
- AlignGVGD thresholds are dependent on the disease. The generic is 0-123456.
- PolyPhen thresholds are also dependent on the disease. The generic is PP-B.
- The best PON tool is dependent on the disease. The generic is SIFT.
- Optimal MSAs are dependent on the disease and PON tool used. No generic could be found.



Uses of framework

- Useful in finding the optimal MSA, PON tool and corresponding thresholds for LARGE data sets of genetic diseases
- Can be used with limited field expertise
- Provides a benchmark against newer MSA or PON tools



Suggested Future Work

- Test the assumption that the best MSA tool gives the better PON prediction
- Further testing of more genetic disease data sets
- Testing of more MSA & PON tools
- PMut thresholds using the confidence values
- Create a fully automated software or workflow
- Create a 'plug-in' module consisting of the optimal MSA for each tool, optimal generic MSA and thresholds – maybe database



Questions ?

Thank You