



Screening for Non-Acceptable Polymorphisms

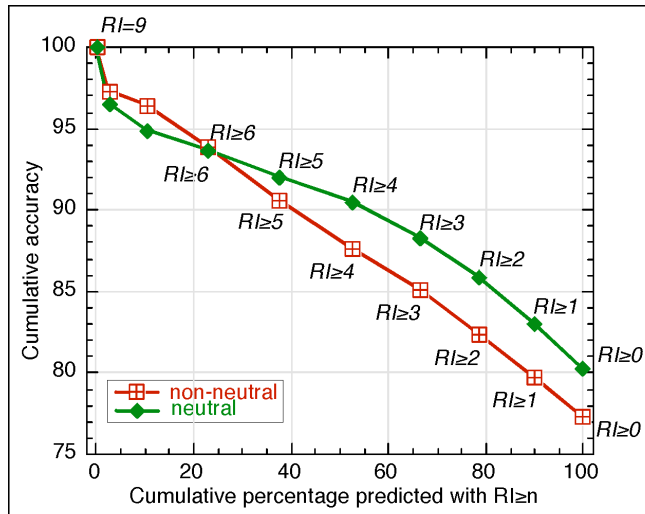
is a neural network based method for the prediction of the functional effects of non-synonymous SNPs. SNAP needs only sequence information as input, but benefits from functional and structural annotations, if available. In a cross-validation test on over 80,000 mutants, SNAP identified 80% of the *non-neutral* (loss and gain of function) substitutions at 77% accuracy and 76% of the *neutral* (functionally unaltered) substitutions at 80% accuracy. This constituted an important improvement over other methods; the improvement rose to over ten percentage points for mutants for which existing methods disagreed (SNAP made a correct prediction when two other well known computational methods (SIFT and PolyPhen) disagreed 1.7 times more often than either of the others). Possibly even more importantly SNAP introduced a well-calibrated measure (reliability index, RI) for the reliability of each prediction. Interestingly, this measure also seems to outline the severity of effect of given substitutions. Thus, the RI allows users to not only focus on the most accurate predictions, but also on the most severe effects.

SNAP features:

- **No discrimination by organism.** SNAP training and testing data sets contained proteins from over 800 different strains and organisms. Although some of these were clearly better represented due to their

extensive presence in research publications, SNAP's overall performance was similar for all.

- **SNAP is more sensitive to severe changes.** SNAP's reliability index (RI) is a good measure of how trustworthy a given prediction is. However, also seems to be acceptable for determining the severity of a given prediction. SNAP was trained only on experimental data of binary nature, i.e. the supervised output was either labeled as non-neutral or as neutral. All the same, the network seems to have learned implicitly to distinguish between more and less severe effects. We did not have enough appropriate data to analyze this question rigorously. We did however have some data of appropriate format (*E. coli* LacI repressor data

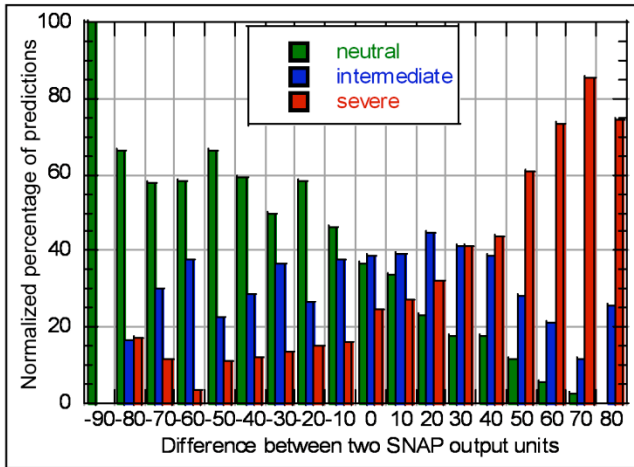


Stronger SNAP predictions are more accurate. The reliability index (RI) effectively estimates the accuracy of a prediction and enables users to focus on more reliable predictions.

set containing 4041 mutants) and a limited grading of severity (neutral / slightly damaging / damaging / severe). SNAP clearly performed better on more mutants with more severe effects and on clearly "more neutral" changes than on those with intermediate effects. This suggested that the difference between the output values did not only

reflect the reliability of predictions, but also the severity of the change. Put differently, more severe effects corresponded to stronger SNAP predictions. The reason why SNAP implicitly learned about the severity of effects was likely of statistical nature: the most severe and most neutral mutations were most consistent in the data set.

- **SNAP could predict gain of function as well as loss of function.** There is a number of nsSNPs in the training data set that are known



Stronger signals for more severe changes. We observe higher RI for more that more severe changes. Samples in the “very slightly damaging” and “slightly damaging” category were combined into a single “intermediate” category. We normalized the predictions in each “severity group” because the samples for “intermediate effects” were significantly under-represented in the experimental data.

to introduce a “gain of function” for a particular protein. Unfortunately, many mutations entail a gain of one function, but a loss, or retention at same levels, of another. Additionally, modifications that lead to loss of function in one protein may very well correlate with a gain of function in another (e.g. increased structural flexibility may suppress or promote function). Given this reality, we chose not to separate out directions of effects of mutation: for the purposes of SNAP a gain of function is treated as a non-neutral sample. However, to illustrate that SNAP can recognize these mutants consider an example of a few nsSNPs of the metalloendopeptidase thermolysin (EC.3.4.24.). A study of this enzyme showed that set of 18 nsSNPs of thermolysin increase its activity (“gain of function”). SNAP correctly identified all of these mutations as non-neutral, with reliability index range 0-5. Although SNAP is already capable of recognizing gain of function changes, it would likely benefit from seeing more of this sort of mutants. However, extensive data of this type is not currently available.

SNAP can accommodate the needs of large-scale experimental scans. The features of SNAP make it particularly useful to researchers who want to scan large experimental data sets. SNAP’s particular strength is making the correct predictions for the least obvious cases (those for which existing methods disagree). For such mutants SNAP was 13-17 percentage points more accurate than other methods. Furthermore, these mutants are generally also the hardest to pinpoint experimentally since their subtle effects are more likely to contribute to a phenotype rather than fully account for it. In a typical experimental scenario in which experimental observations are likely to have already been subjected to various analysis tools, this improvement is likely to be extremely relevant and significant.

Gene	Org.	nsSNP	Phenotype	Func	Predict (RI)
HXK4	Human	S131P	Diabetes Mellitus	↑	Non-neut (2)
PAX6	Human	G64V	Cataract	↓	Non-neut (5)
HBL1	Rice	H74L	Reduced O2 affinity	↓	Non-neut (6)
MC4R	Human	T112M	Wild-type	↓	Non-neut (1)
HXK4	Human	M107T	Wild-type	?	Neutral (0)
CFTR	Human	P1013L	Cystic Fibrosis	?	Non-neut (5)
NKX25	Human	A127G	Atrial septal defect	?	Neutral (6)
P53	Human	R337H	Carcinoma	↔	Non-neut (3)
OASA1	Rice	D323N	Reduced Trp sensitivity	?	Neutral (2)

Examples of SNAP predictions for mutants associated with a known phenotype and/or known functional change. In function (↑ represent increase or decrease of function, (?) represents unknown effect, and (□) means that one protein function was unchanged and others are unknown. In the table the green fields correspond to full agreement between prediction, phenotype, and functional annotation. The yellow fields represent the twilight zone of biology where the experimental functional annotation either doesn't exist or doesn't agree with a disease phenotype. The red fields display the possible disagreements between predictions and biology: in the first cases the mutant responsible for the observed residue substitution is likely in linkage disequilibrium with the phenotype-causing mutation (hence disease association, but neutral prediction). In the second case, the function that was measured is likely not the function that is affected and leads to the observed phenotype.