Precision Diagnosis for Rare Diseases

2019-11-08

Precise Diagnosis through Sensitive Computational Approach: One Gene at a Time

Prediction of the effects of genetic variants in the context of health and disease has been a significant effort for clinicians, biologists, and computational scientists. Despite this extensive effort, the performance of such tools is not at the desired level. Therefore, clinicians are hesitant about using them in practice.

Let's work on a potential scenario.

A medical doctor has a patient. The doctor suspected of a rare genetic disease, and she sequenced her patient's DNA. Doktor received the variants and found two heterozygous mutations in the gene that is associated with the disease. She would like to know the significance of these variants. What are her options:

1) **Cataloged variant.** If the variants of interest have been studied before, she can generate here conclusions derived from previous studies with confidence. The case is closed.

2) **Obvious results.** The variants under investigation might have apparent deleterious effects on the protein function. For instance, if the variant is in the conserved spliced site or if it introduces an early stop codon, then the consequence would be "deleterious."

3) **Frequent allele.** If the variants of interest are seen in the human population at a relatively high percentage, which cannot cause the suspected rare disease, then the doctor can easily categorize the variants as "benign."

4) **Novel mutations.** If the mutation that is under investigation is novel, meaning that it has never been studied and observed, then we are left with functional studies or computational prediction. As functional tests are impractical in many of the research groups, they often choose to use computational tools.

How do currently available tools predict the effect of variations?

There are numerous tools available to be used to predict the nucleotide or amino acid variations. For the protein-coding genes, protein sequence is often used rather than DNA. Most of the tools use Evolutionary Information as a sole or partial component in their algorithms. Some algorithms also use population data, protein structure, etc. as complementing data sets. Being the primary source of information, evolution has a great significance in predicting the deleterious effect of a variation. The idea is simple: if an amino acid has been conserved for millions of years of evolution, then it cannot be "touched." Any change in that position will be harmful to the protein function and, therefore, health. For this reason, evolutionary information is essential to be generated meticulously.

Currently, the simplest way of building an evolutionary history of a gene starts with collecting homologs (genes/proteins which are derived from a common ancestor). After some filtering, multiple sequence alignment is generated. With multiple sequence alignment, conserved and variable residues are revealed. After

Although the pipeline sounds reasonable. is it the entire evolutionary information that we can retrieve from the genomic data sets derived from a range of eukaryotic organisms. **We believe we can build more meaningful evolutionary histories.**

Problems in the current state of the approaches

Evolutionary dependent variants

Most algorithms do not take phylogenetic information into account when building their models. We, however, think that it is crucial. Here is why:

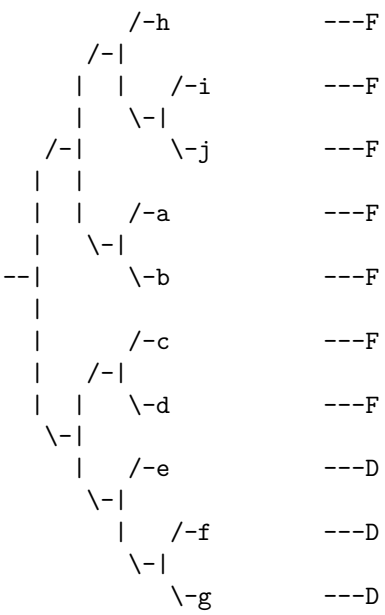
If in the multiple sequence alignment, the position of interest is fully conserved (e.g., 100% Phenylalanine [F]), then it is inferred that the position cannot be "touched." If the mutation is on that position, it likely distrupts protein function and it is associated with the disease.

If in MSA, the position is changed only for one time (e.g., 9/10 is F, and 1/10 is K), then it is still not possible to talk about the neutrality of the F->K substitution.

However, if the position has partly conserved and substituting amino acid is observed in other organisms more than once (eg, 7/10 F and 3/10 K), then algorithms can predict that F->K substitution is "harmless" as it has been observed in evolution more than one time.

Is that so?

Imagine that 3 F->K substitution occurred in the common ancestor of these three organisms, such as the below phylogenetic tree.



As you can see in this case, the F->K substitutions occurred in the common ancestor of all organisms having K at the position of interest. Therefore, this is a single evolutionary event. When this substitution occurred, another compensatory change could have also occurred. As is, this scenario is not different than the one we mentioned above where we have single K residue in the MSA.

To sum up, rather than counting the number of residues for substitution, we should be counting the number of **evolutionary events.** Obtaining evolutionary events is only possible through phylogenetic analysis.

Paralogs

Gene duplication is the primary mechanism to invent a new protein. When a gene is duplicated, there are three possible scenarios:

i) One of the genes accumulate mutations and become a pseudogene: **Non-functionalization**.
ii) Two genes accumulate a different set of variations, and both perform a part of the original function and usually complete each other's role: **Sub-functionalization**.
iii) One of the genes preserves its purpose, and the other one accumulates mutations and gains a new function: **Neo-functionalization**.

For none of the cases, after gene duplication, genes (or proteins) have an identical function. Therefore paralogs should be considered separately.

We study the evolutionary history of a gene carefully to identify all evolutionary events in the family. We reveal orthologs, paralogs as well as functional orthologs through phylogenetic analyses. At the end of the investigation, we build our gene-specific algorithm based on a clean set of sequences generating an MSA.

We improve the sensitivity of predicting disease-causing mutations

We implemented our approach to accurately predict disease-causing mutations for Niemann Pick Disease Type C (NPC). We studied the evolutionary history of the NPC1 gene, which is responsible for >90% of the NPC cases. Here is the abstract of our work and the link for the original paper published in Genetics in Medicine.

Predicting the phenotypic effects of mutations has become an essential application in clinical genetic diagnostics. Computational tools evaluate the behavior of the variant over evolutionary time and assume that variations seen during evolution are probably benign in humans. However, current tools do not take into account orthologous/paralogous relationships. Paralogs have dramatically different roles in Mendelian diseases. For example, whereas inactivating mutations in the NPC1 gene cause the neurodegenerative disorder Niemann-Pick C, inactivating mutations in its paralog NPC1L1 are not disease-causing and, moreover, are implicated in protection from coronary heart disease.

We identified significant events in NPC1 evolution and revealed and compared orthologs and paralogs of the human NPC1 gene through phylogenetic and protein sequence analyses. We predicted whether an amino acid substitution affects protein function by reducing the organism's fitness.

Removing the paralogs and distant homologs improved the overall performance of categorizing disease-causing and neutral amino acid substitutions.

The results show that a thorough evolutionary analysis followed by the identification of orthologs improves the accuracy in predicting disease-causing missense mutations. We anticipate that this approach will be used as a reference in the interpretation of variants in other genetic diseases as well.