12-2015

Applications of Evolutionary Bioinformatics in Basic and Biomedical Research

Ogun Adebali
University of Tennessee - Knoxville, oadebali@vols.utk.edu

Recommended Citation

Adebali, Ogun, "Applications of Evolutionary Bioinformatics in Basic and Biomedical Research. " PhD diss., University of Tennessee, 2015.
https://trace.tennessee.edu/utk_graddiss/3540

To the Graduate Council:

I am submitting herewith a dissertation written by Ogun Adebali entitled "Applications of Evolutionary Bioinformatics in Basic and Biomedical Research." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Igor B. Jouline, Major Professor

We have read this dissertation and recommend its acceptance:

Albrecht von Arnim, Jerome Baudry, Elias Fernandez

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Applications of Evolutionary Bioinformatics in Basic and Biomedical Research

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Ogun Adebali

December 2015

DEDICATION

To my partner.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my mentor Dr. Igor B. Jouline for his supervision, trust and support throughout my research. I cannot thank him enough for expanding my scientific view. I appreciate my dissertation committee members, Dr. Albrecht von Arnim, Dr. Jerome Baudry and Dr. Elias Fernandez for their discussions and critiques.

I am grateful to my lab mates: Dr. Davi Ortega, Dr. Luke Ulrich, Amit Upadhyay and Dr. Aaron Fleetwood for their teachings, discussions and contributions. Great friends, Reka Kelemen and Caroline Rempe also deserve thanks for giving me important feedbacks on my writings. I would also like to thank Gerald Ragghianti who is the administrator of Newton High-Performance Cluster at the University of Tennessee for his advices on technical issues.

I would like to acknowledge the staff and faculty of the Genome Science and Technology (GST) program for maintaining such a great graduate school. I wouldn’t be able to present this work today without the support and funding from GST, the Microbiology Department, Oak Ridge National Laboratory and National Institute of Health.

Special thanks go to my family members who had to be an ocean away from us for a considerable amount of time.

Finally, this dissertation is dedicated to my wife, Yasemin Adebali, who has been with me all the time since the beginning of this journey.

ABSTRACT

With the revolutionary progress in sequencing technologies, computational biology emerged as a game-changing field which is applied in understanding molecular events of life for not only complementary but also exploratory purposes. Bioinformatics resources and tools significantly help in data generation, organization and analysis. However, there is still a need for developing new approaches built based on a biologist's point of view. In protein bioinformatics, there are several fundamental problems such as (i) determining protein function; (ii) identifying protein-protein interactions; (iii) predicting the effect of amino acid variants. Here, I present three chapters addressing these problems from an evolutionary perspective. Firstly, I describe a novel search pipeline for protein domain identification. The algorithm chain provides sensitive domain assignments with the highest possible specificity. Secondly, I present a tool enabling large-scale visualization of presences and absences of proteins in hierarchically clustered genomes. This tool visualizes multi-layer information of any kind of genome-linked data with a special focus on domain architectures, enabling identification of coevolving domains/proteins, which can eventually help in identifying functionally interacting proteins. And finally, I propose an approach for distinguishing between benign and damaging missense mutations in a human disease by establishing the precise evolutionary history of the associated gene. This part introduces new criteria on how to determine functional orthologs via phylogenetic analysis. All three parts use comparative genomics and/or sequence analyses. Taken together, this study addresses important problems in protein bioinformatics and as a whole it can be utilized to describe proteins by their domains, coevolving partners and functionally important residues.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ATTACHMENTS

CHAPTER 1. Introduction

The genomics era was started in 1995 by a group led by Craig Venter who published the complete DNA (genome) sequence of *Haemophilus influenza* (Fleischmann et al. 1995; Cristianini and Hahn 2006). It was the first sequenced genome of a free-living organism. Venter's method of genome assembly, shotgun sequencing, opened the doors for inevitable genomic data accumulation. In the following year, the first eukaryotic genome (*Saccharomyces cerevisiae*) was published (Goffeau et al. 1996). Year by year, DNA sequencing became easier and cheaper. After only 5 years, the human genome was sequenced (Venter et al. 2001). Today, researchers (from the United States and China in parallel) aim to sequence 1 million human genomes in the next few years (Stephens et al. 2015).

With the revolutionary developments in sequencing technologies and instrumentation, genomic data has been accumulating exponentially **(Figure 1-1)**. So far, the amount of genomic data has doubled every 7 months since 2008 which has significantly beaten the curve of Moore's law (doubles every 18 months) (Stephens et al. 2015). As of August 2015, the number of sequenced (partially/fully) genomes in the NCBI genome database approached 50,000. Though gaining more information on living beings, including humans, sounds like a great advancement, the ability to understand the data remains a growing challenge. Since generated raw sequence data alone is not useful to infer information, unprocessed sequence data needs to be compiled and converted into a format that addresses certain questions of interest. This data-processing is performed with specialized tools through scientific measurement methods. Bioinformatics, which is partly the field of designing tools and approaches for retrieving, storing, organizing, visualizing and analyzing biological data, is vital for scientists desiring to know about their

DNA/protein sequence of interest (Luscombe et al. 2001). New knowledge derived from living beings can be obtained by using the services offered by the bioinformatics field. However, making sense out of large series of letters (DNA or protein sequences) is challenging. As the interpretation of large data depends on computational capabilities, the fields of computational biology and bioinformatics are continuously expanding since the beginning of the exponential growth of biological data. Research in these fields attracted attention due to promising results for prognosis, diagnosis and treatment of diseases (Mount 2001).

There are three aims of bioinformatics (Kumar and Dudley 2007). The first aim is to develop systems to store data, so that it is accessible to researchers. Because automated algorithms are erroneous, manual curation is important (Howe et al. 2008). However, manual curation is a limiting step and tends to fall behind of the growth of genomic data. Though automated systems are still not at the desired level of precision, as our understanding of molecular biology expands, the quality of these resources increases. For instance, the partially manually-curated protein database RefSeq has not shown a steeper rate of increase especially in the last few years **(Figure 1-2)**. The second aim is to develop tools, methods and resources helping biologists to convert data to knowledge. Because the majority of researchers conducting experimental work are not data scientists, user-friendly tools are needed (Kumar and Dudley 2007). The increasing rate of genomic data has surpassed computational developments. Therefore, algorithms that are specifically designed for genomic data should be developed and implemented to maintain an efficient computation for a desired task. Bioinformatics tools and resources can be diverse. While some of them are general, most bioinformatics tools are designed

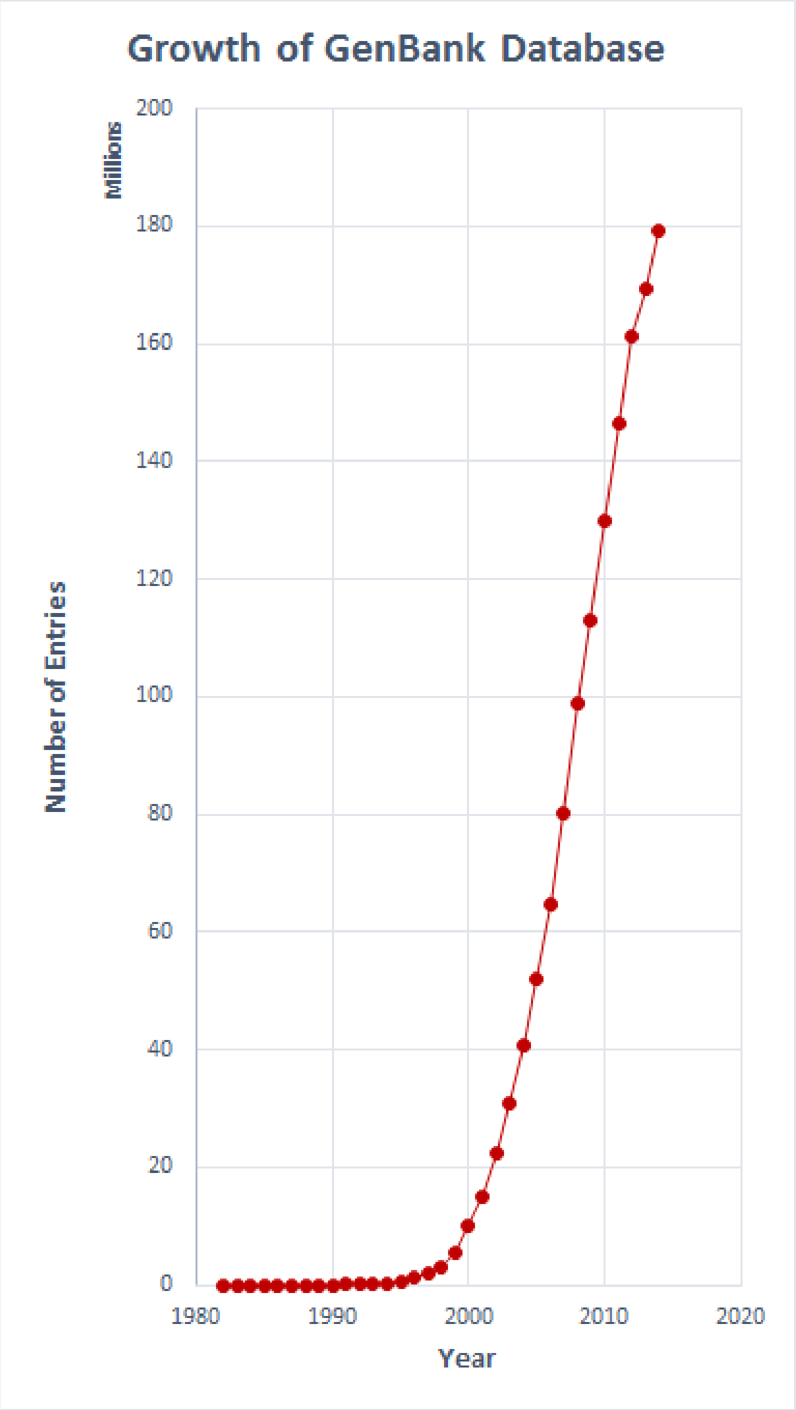


Figure 1-1 Genomic data growth in GenBank.

The data retrieved from GenBank.

to address specific questions of interest. Finally, the third aim is to apply these resources in data analysis in order to infer biologically meaningful results. This dissertation serves promote to the last two aims of bioinformatics with attempts to understand proteins from an evolutionary perspective. As a subfield of biology, evolutionary bioinformatics focuses on understanding mutative events at the levels ranging from molecules to populations, by applying computational methods. This work concentrates on protein evolution at not only a molecular but also a genomic level.

1.1 Biology overview

1.1.1 Central dogma of molecular biology: DNA to protein

The genome of a free-living organism is composed of DNA (deoxyribonucleic acid) which carries the heritable information. DNA is in double-helix conformation (Watson and Crick 1953) and each strand is composed of a linear sequence of deoxyribonucleotides: adenine (A), thymine (T), guanine (G) and cytosine(C). A and G (purines) pair with T and C (pyrimidines), respectively, through hydrogen bonding (Donohue and Trueblood 1960). Due to this complementarity the sequence of a single strand is sufficient to determine the sequence of the complementary strand.

Originally described by Francis Crick in 1958, the central dogma of molecular biology explains the flow of information residing on biopolymers of DNA, RNA (ribonucleic acid) and protein in a residue-by-residue fashion **(Figure 1-3)** (Crick 1958; Crick 1970). Although the rule oversimplifies complex biological information transfer, it is useful to understand how protein is synthesized by the information coded in DNA (Biro 2004).

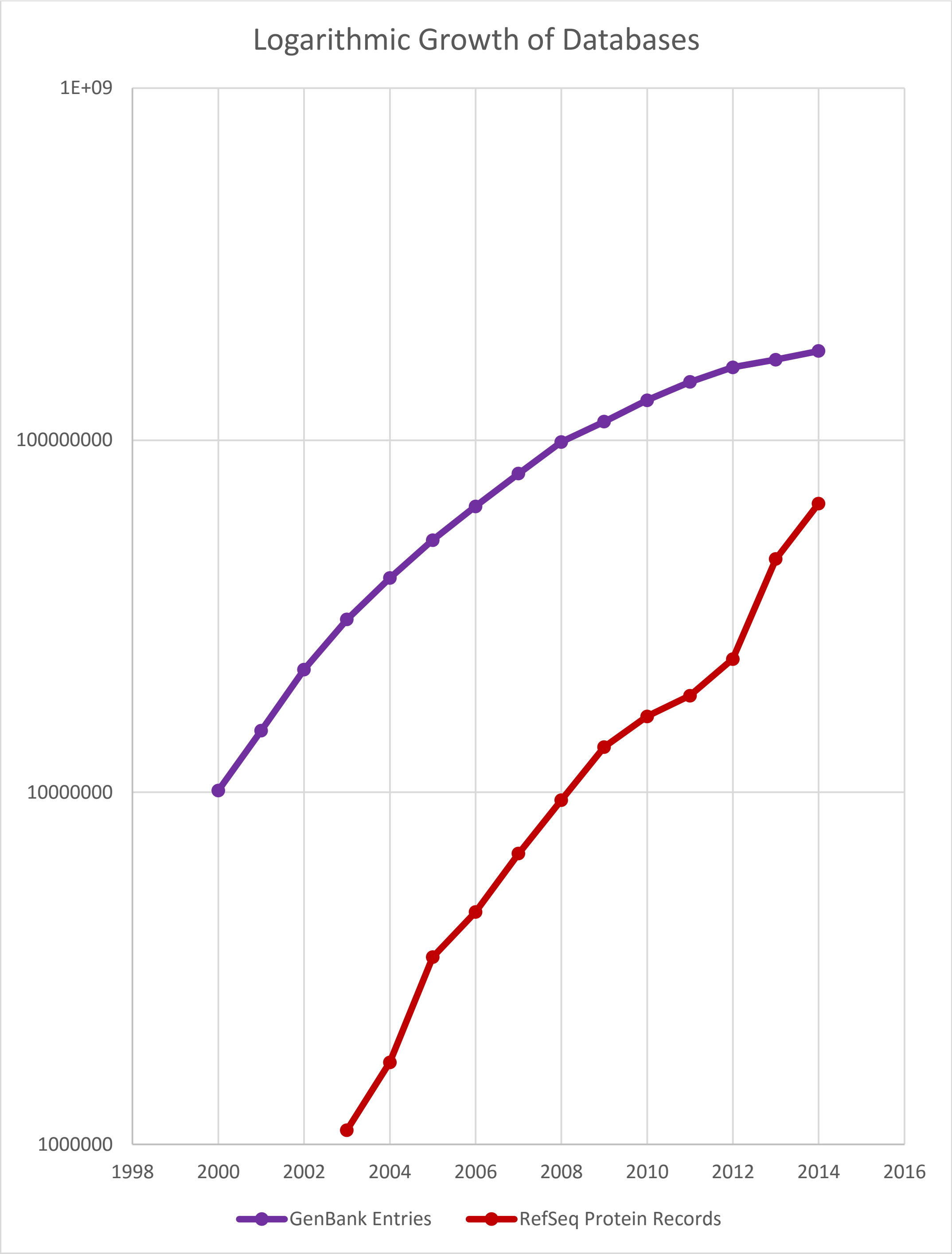


Figure 1-2 The logarithmic growth of GenBank and RefSeq databases.

As it is manually curated, the RefSeq database is harder to construct. However, in the few last years, there was an increase in the number of records in RefSeq. This rate of increase exceeded the GenBank rate.

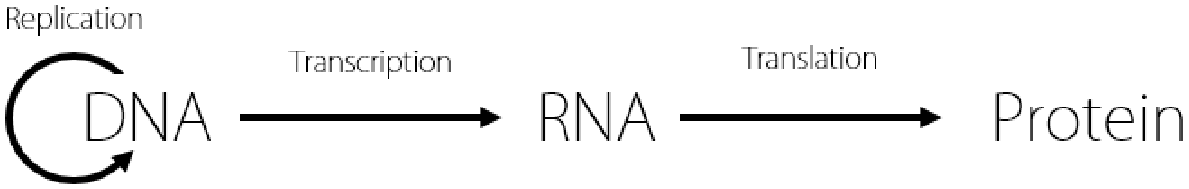


Figure 1-3 Central dogma of biology.

In protein synthesis, the first step is to construct messenger RNA (mRNA) from a genomic DNA template (complementary strand) in a process called transcription. RNA polymerase starts mRNA synthesis by binding an upstream region (promoter) of the gene. This binding is usually aided by proteins called transcription factors. RNA polymerase reads a single strand of DNA and synthesizes RNA in 5' → 3' direction. The mRNA sequence is identical to the coding strand (complementary to the template strand) except for uracil (U) in place of thymine (T).

In eukaryotes, most genes include regions that do not code for protein. These regions are called introns. Right after eukaryotic transcription the mRNA (often called pre-mRNA) contains both coding regions (exons and introns). However, the pre-mRNA is processed and intron regions are excluded by joining exons. This process is called RNA splicing **(Figure 1-4)**. During splicing, some of the exons can be skipped and various combinations of exons (while keeping the order intact) result in unique mature mRNA products. As the name alternative splicing implies, this process eventually results in different protein products (isoforms) coded by the same gene. Eukaryotic transcription and splicing take place in the nucleus and mature RNA are then transferred to the cytoplasm.

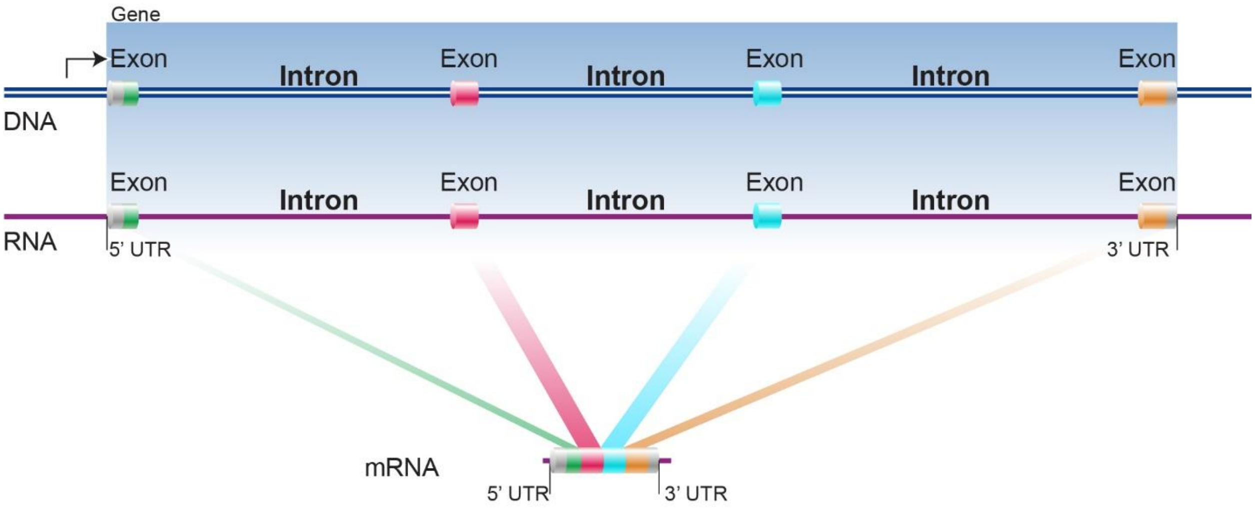


Figure 1-4 mRNA splicing.

Copied from National Human Genome Research Institute.

In eukaryotes, after mature mRNA is produced, mRNA is transported from the nucleus to the cytoplasm. Here, a complex protein/RNA structure, the ribosome, facilities information transfer from mRNA to protein. This process is called translation. In eukaryotes, first, the ribosome binds to the untranslated region (UTR) of mRNA at the 5' end. Then, the mRNA is scanned in a 5' → 3' direction to search for the start signature of three letters: AUG (start codon). mRNA in a 5' → 3' direction corresponds to an N terminus to C terminus directionality in proteins. The mRNA sequence is translated into an amino acid sequence in protein synthesis. Each 3-letter nucleotide block of RNA (codon), codes for a corresponding amino acid. The 4 nucleotides, A, G, U and C can generate 64 (4^3) different codons **(Figure 1-5)**. 3 codons (UAG, UAA and UGA) do not code for amino acids, as they are stop codons. Protein synthesis stops when a ribosome encounters any stop codon.

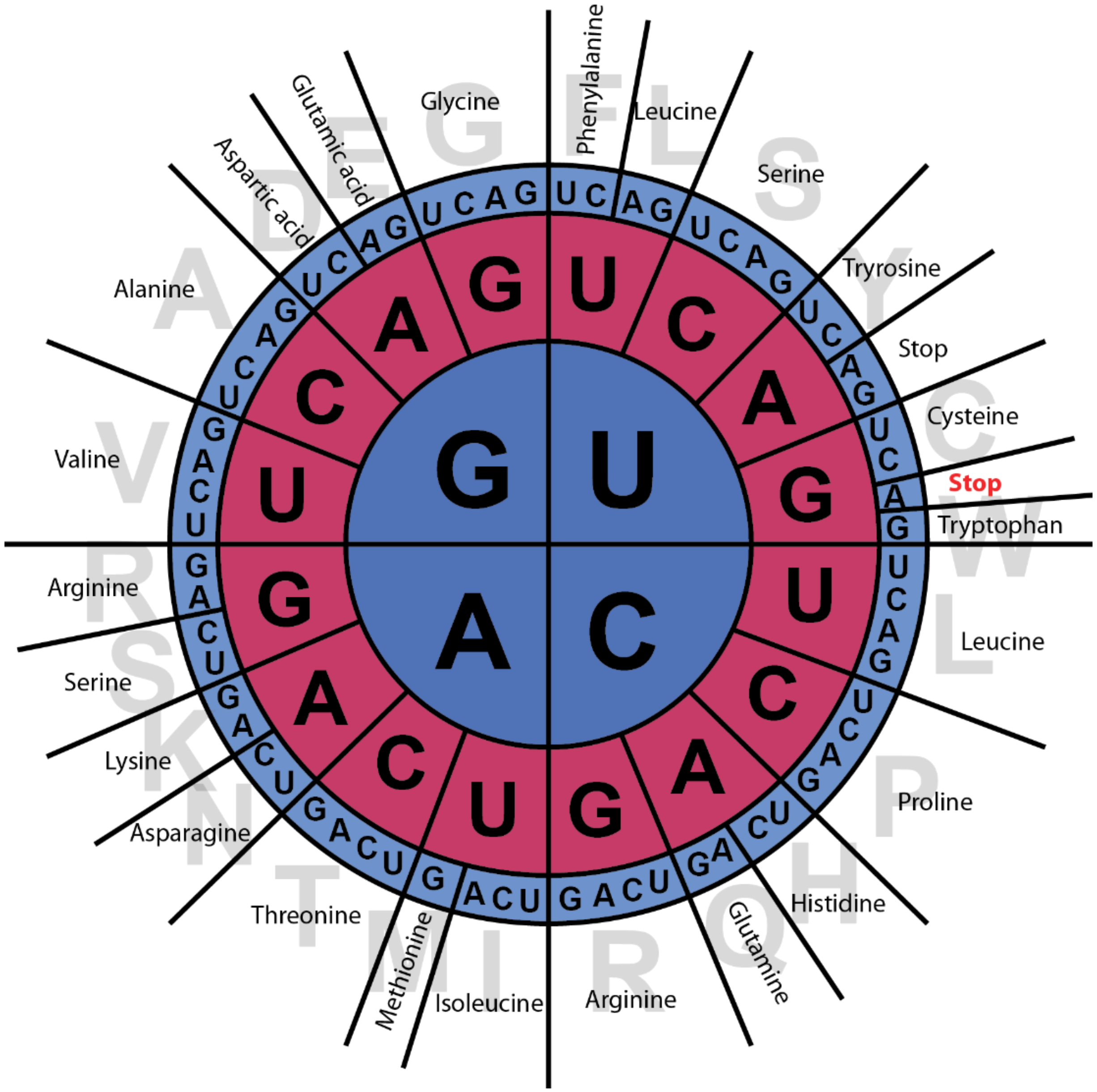


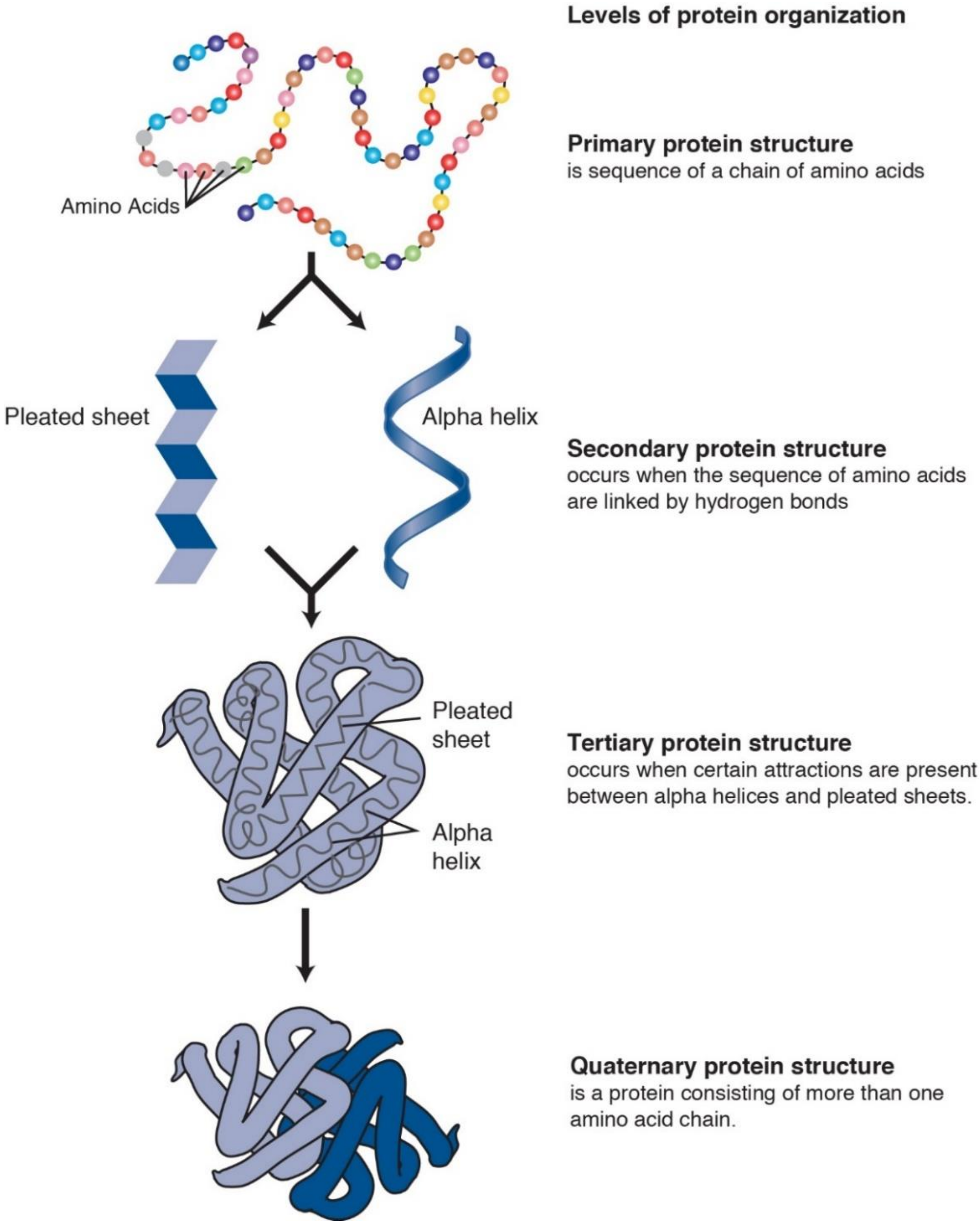
Figure 1-5 Codon wheel and single letter codes for amino acids.

After translation, a polypeptide chain (primary protein structure) is folded into a three-dimensional structure **(Figure 1-6)**. First, certain segments of protein are folded into general structural elements (secondary structures: alpha helix or beta sheet). Then, physical attractions between secondary structural elements result in more folding which eventually results in tertiary protein structure. Tertiary structures of the same protein can come together and form a complex (oligomer). This complex is called quaternary protein structure. Protein folding is primarily determined by the amino acid sequence itself. External factors such as solvent properties, temperature, and other aiding proteins also play important roles in folding. Because the structure of a protein is directly related to the function, deficiency in folding causes function disruption.

Three-dimensional protein structures are composed of one or more functional units called domains **(Figure 1-7)**. Domains are minimal functional and structural elements of proteins that can fold autonomously and evolve independently from the rest of protein as they are found in various domain arrangements in proteins.

1.1.2 Mutations

A mechanism exists that can make permanent and heritable changes to DNA, thus allowing genetic diversity and speciation. These changes can be due to environmental factors, viruses and transposable elements, and erroneous replication where DNA polymerases make errors. DNA polymerase makes errors at a certain rate and most of them are corrected by molecular proof-reading mechanisms (Reha-Krantz 2010). Uncorrected DNA changes are called mutations. DNA mutation and recombination are the major reasons for genetic diversity in a population. These molecular events are also



Levels of protein organization

Primary protein structure
is sequence of a chain of amino acids

Secondary protein structure
occurs when the sequence of amino acids are linked by hydrogen bonds

Tertiary protein structure
occurs when certain attractions are present between alpha helices and pleated sheets.

Quaternary protein structure
is a protein consisting of more than one amino acid chain.

Figure 1-6 Protein folding - from polypeptide chain to the three-dimensional structure.

Copied from National Human Genome Research Institute.

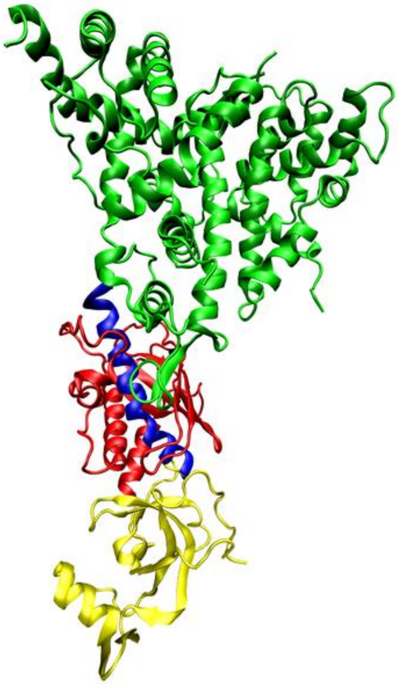


Figure 1-7 Example of a multi-domain protein.

A Dicer protein contains three domains and a linker. Green: RNase III domain, yellow: PAZ domain, red: platform domain. The linker is shown as blue.

responsible for many genetic disorders. Mutations can be classified in two general categories: (i) substitution and (ii) indel (insertion or deletion).

Mutations accumulate in non-coding regions at an incomparably high rate relative to coding regions. Because of their relatively low importance in protein function, there is no eliminating selection pressure on these regions. However, some regions are critical in protein synthesis such as promoters, enhancers and silencers which play roles in protein expression. These regions are less prone to accumulate mutations compared to other non-coding regions because of their vital roles that are conserved for survival. Mutations in coding regions are less tolerable and thus less frequently observed than non-coding variations. There are several types of substitution mutations within a coding region **(Figure 1-8)**. First, a single nucleotide mutation resulting in a different codon that still encodes the original amino acid is called a *synonymous* or *silent* mutation. Synonymous mutations are mostly tolerable because the change doesn't affect protein sequence and thus structure and function. Secondly, *missense* mutations cause single amino acid replacement on the protein and can affect protein folding, catalytic activity, and interaction with other molecules. Some missense mutations can be neutral and have no effect on protein function while others can slightly or severly change protein activity. Finally, *non-sense* mutations change amino acid-encoding (sense) codons to stop (nonsense) codons where protein synthesis is forced to terminate early. The truncated polypeptide chains are usually non-functional. Beside substitutions, indels are also important in protein function. DNA polymerases might skip reading one or more nucleotides on the template, which causes deletion in the newly synthesized DNA. On the contrary, these enzymes may also add extra nucleotides while copying DNA, resulting in insertion. Indels are more likely to affect protein function if the length of indel is not a multiple of three (size of codons). This changes the codons read during translation, resulting in a meaningless protein sequence. This type of indel mutation causes a shift of the open reading frame called a "frame-shift" mutation.

In diploid organisms such as humans, every gene has two copies (alleles), one on each pair of chromosomes. Alleles can be identical or different in terms of sequence. For a specific trait, if alleles result in the same observation (phenotype) the genetic condition of alleles is called homozygous. If alleles yield different proteins for a certain trait then they are called heterozygous.

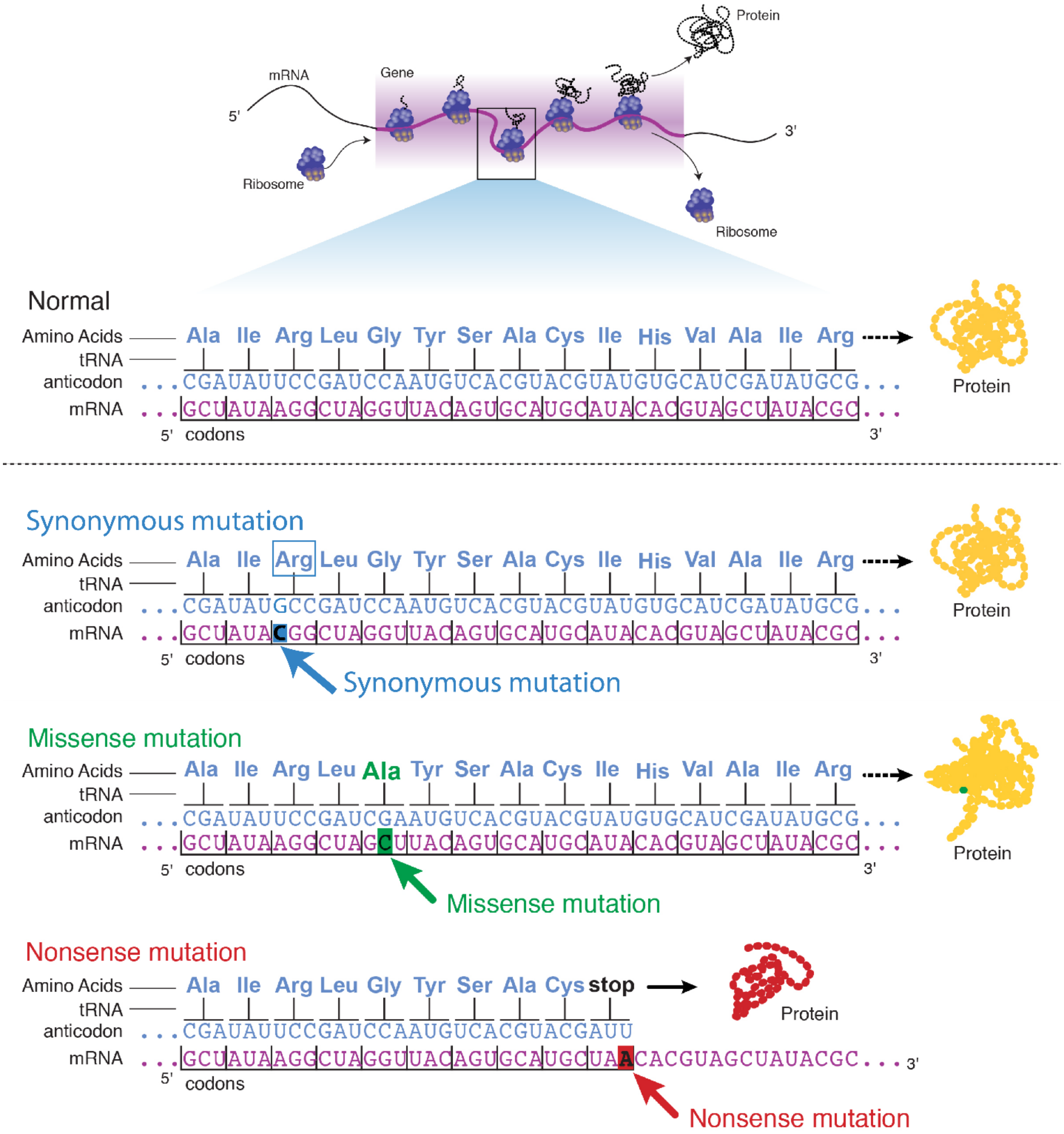

Protein
mRNA
Gene
5'
3'
Ribosome
Ribosome
Normal
Amino Acids
tRNA
anticodon
mRNA
Ala Ile Arg Leu Gly Tyr Ser Ala Cys Ile His Val Ala Ile Arg
. . . CGAUAUUCCGAUCCAAUGUCACGUACGUAUGUGCAUCGAUAUGCG . . .
. . . GCU AUA AGG CUA GGU UAC AGU GCA UGC AUA CAC GUA GCU AUA CGC . . .
5' codons
3'
Protein
Synonymous mutation
Amino Acids
tRNA
anticodon
mRNA
Ala Ile Arg Leu Gly Tyr Ser Ala Cys Ile His Val Ala Ile Arg
. . . CGAUAUGCCGAUCCAAUGUCACGUACGUAUGUGCAUCGAUAUGCG . . .
. . . GCU AUA CGG CUA GGU UAC AGU GCA UGC AUA CAC GUA GCU AUA CGC . . .
5' codons
3'
Synonymous mutation
Protein
Missense mutation
Amino Acids
tRNA
anticodon
mRNA
Ala Ile Arg Leu Ala Tyr Ser Ala Cys Ile His Val Ala Ile Arg
. . . CGAUAUUCCGAUCGAAUGUCACGUACGUAUGUGCAUCGAUAUGCG . . .
. . . GCU AUA AGG CUA GCU UAC AGU GCA UGC AUA CAC GUA GCU AUA CGC . . .
5' codons
3'
Missense mutation
Protein
Nonsense mutation
Amino Acids
tRNA
anticodon
mRNA
Ala Ile Arg Leu Gly Tyr Ser Ala Cys stop
. . . CGAUAUUCCGAUCCAAUGUCACGUACGAUU
. . . GCU AUA AGG CUA GGU UAC AGU GCA UGC UAA CACGUAGCUAUACGC . . . 3'
5' codons
Protein
Nonsense mutation

Figure 1-8 Types of single point mutations.

Adapted from National Human Genome Research Institute.

If one of the alleles in a diploid organism causes a defective protein due to a mutation while the other allele codes for a functional protein, there are three possible consequences: (i) protein function is complemented by the protein coded by the wild-type allele (functional allele is dominant to the defective one), (ii) the deficiency in one allele results in insufficient amount of fully functional protein, so the overall function remains deficient (defective allele is dominant), (iii) the lack of protein amount partially affects protein function (partial dominance or haploinsufficiency).

There are rare and common variants in a population. Simple Mendelian diseases and novel genetic disorders are mostly caused by rare variants. Genetic variants in a population are called polymorphisms. Although polymorphisms are usually benign, some of them have been found to be associated with diseases (Satake et al. 2009), especially certain combinations of single nucleotide polymorphisms (SNPs) (De Gobbi et al. 2006).

1.1.3 Molecular evolution

Gene duplication is the major mechanism in the formation of new functions in evolution (Ohno 2013). Gene duplication occurs via several molecular events such as unequal crossing-over, DNA polymerase slippage, retrotransposable elements and non-disjunction during meiosis (Ohno 2013). When a gene is copied, an identical gene can keep the function if it was necessary for the organism and fixed in the population. However, the other copy remains redundant at first and prone to change and accumulation of mutations. Over generations, this copy may diverge and result in three possible consequences: It is (i) lost because the sequence doesn’t constitute a meaningful biomolecule; (ii) neo-functionalized and gains a new function; (iii) subfunctionalized and takes on a subset of functions of the original gene (Lynch and Conery 2000).

Most gene families are enlarged due to gene duplication. Genes and proteins that share a common ancestor are called homologs. If the homology is achieved through gene duplication within a current or ancestral species, homologous genes are called paralogs. Homologous genes in different species that evolved by a speciation event are termed as orthologs. Orthologs are more likely to conserve their function while paralogs are not anticipated to keep the identical function (Taylor and Raes 2004).

Natural selection is a key principle in evolution and well reflected at the molecular level. Genes that are not vital for survival keep changing until they are either lost or modified enough to gain a new function. We see the same trend in amino acid sequences. Some residues are critical and cannot be replaced by any other amino acid while others are unimportant for function and structure, so they can be replaced by other amino acids. Amino acids sharing a common physicochemical property which is needed in a particular sequence position of a protein are likely to substitute each other **(Figure 1-9)**.

1.2 Problem overview and motivation

Despite a steep increase in genomic data accumulation, there is still a vast amount that is unknown about the molecular biology of proteins. Particularly, functions of proteins are questions of interest. In order to establish the precise function of a protein, its interactions at a cellular level and physical properties at a molecular level must be determined. Although wet-lab experiments are informative and essential to produce basal data from which computational predictions can be derived, they are expensive and most of them

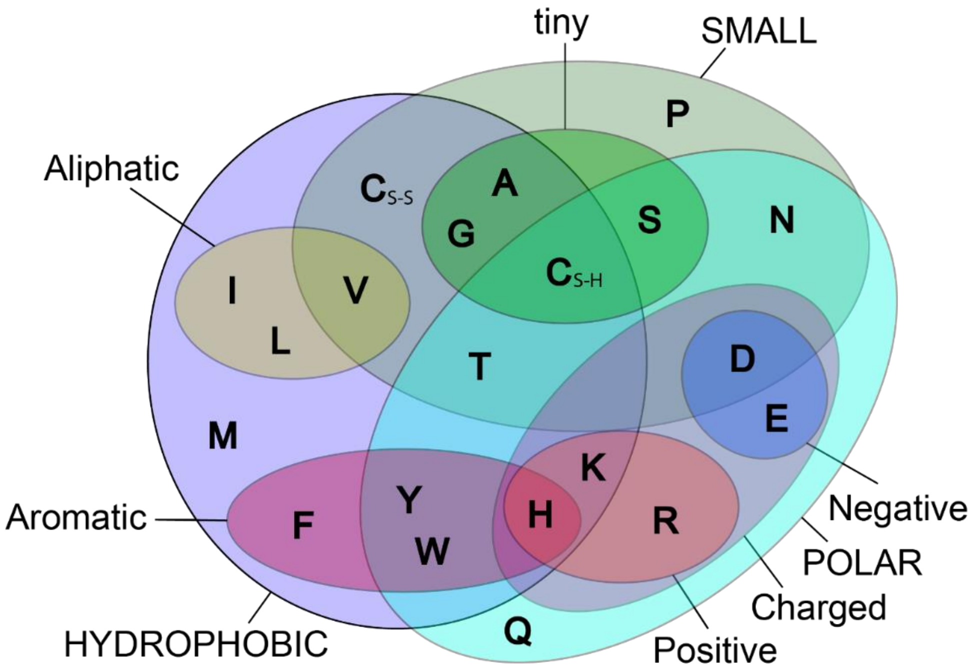


Figure 1-9 Physiochemical properties of amino acids.

are not practical to be applied on a large scale. For this reason, there has been a tendency to computationally predict protein function. In order to determine the general function of a protein, it should first be described based on the sequence content. Computational prediction of protein function at the cellular level depends on two major approaches; (i) homology-based; (ii) context-based. A homology-based method uses a comparative approach and finds similar proteins whose functions are known. If the similarity between “unknown” and “known” proteins is significant, then a prediction of the protein function can be made. However, for non-significant similarity there is still a good chance of predicting protein function. Remote homologs share low sequence identity, but they carry signatures of general properties of protein domains. For this reason, protein domains are computationally predicted. Determining domains in a protein gives insight about the overall function and potential interacting partners. However, determining protein domains is a challenge. There are specific and sensitive tools available. The problem is that specific tools miss domains, while sensitive tools yield an undesired number of false positives. In addition, sensitive tools are computationally intensive and, therefore, they are not applicable for large-scale analyses. Consequently, there is a need for an optimized approach in domain prediction to reach the highest possible confidence while increasing sensitivity.

Content-based protein function prediction is computationally performed by three methods: (i) Co-location; (ii) gene fusion and (iii) genomic co-occurrence (Aravind 2000). Genomic co-location is highly successful in prokaryotes because of the presence of gene clusters (operons) in which genes are regulated together. However, this method is not applicable in eukaryotes because of dispersed locations of interacting genes. Gene fusion is another indicator of interacting genes and proteins. If two or more genes are fused and yield a single protein product in a single organism, it indicates that these proteins are likely to interact in other systems when they are independently located in the genome. Finally, genomic co-occurrence is another indicator of potentially interacting proteins. If two genes/proteins evolve together (they are lost or kept in the genome together), it is likely that they are interacting at least functionally if not physically. This interaction information provides an important understanding about protein function. However, neither gene-fusion nor co-occurrence patterns is straightforward to discover. A major challenge in these analyses is the uncertainty about the absence of biomolecules. Proving true negative in genomic context is not an easy task. However, consistent observation of independent co-absences of genes/proteins would add a confidence. Hence, there is a need for a platform to visualize co-occurrence patterns of protein and domain families to reveal gene fusions and interacting partners.

Even if the overall function of a protein is established, predicting the molecular function of each residue is as critically important. Establishing function of specific residues is important for health and disease in terms of drug design, personalized medicine, and variant outcome prediction. Although knowing specific functions for each amino acid in a protein remains a big challenge, it is possible to weigh their importance in the overall function by evaluating their evolutionary history. By comparing homologous proteins from different species, it is possible to observe which amino acid was conserved and likely to be important and which position was less conserved so that it was replaced by other amino acids in its evolution. There are automated tools performing this task; finding similar protein sequences and comparing each position to assess their weight by evolution. This is a commonly used approach for also predicting damaging and benign mutations for human health and disease. However, the automated tools do not discriminate between orthologs and paralogs. Orthologs are expected to keep the function while paralogs are divergent copies with potentially different functions despite being homologs. Usually, only one of the paralogs is associated with a disease in the same organism for Mendelian disorders. Furthermore, automated algorithms usually result in an approximation when they build an evolutionary history of a gene. This generalization can be disadvantageous when dealing with proteins with distinct evolutionary histories. For these reasons, there is a need for an approach to establish correct evolutionary parameters and separate orthologs from paralogs in evaluating the evolutionary importance of each amino acid position of a protein.

1.3 Scope of dissertation

This dissertation focuses on the molecular understanding of proteins from an evolutionary aspect. The presented work attempts to address three questions: (i) What is the function of a protein; (ii) Which proteins functionally interact with each other; (iii) What is the impact of each amino acid for protein function.

Chapter One briefly introduces basic biology and explain the problems which were addressed in this work. Chapter Two describes the tools, resources and concepts that are important for understanding the rest of the dissertation. Additionally, Chapter Two also discusses the status quo of protein domain exploration and missense mutation outcome prediction. Chapter Three covers an approach for domain identification along with a web-based tool, CDvist. The rationale, approach, and algorithm are briefly explained in this chapter. In Chapter Four, another tool, Aquerium, for phylogenetic profiling visualization is introduced. The web-server also offers a resource for protein domain architecture investigation. In Chapter Five, a rationale for the importance of revealing precise evolutionary history of proteins in health and disease is established with a case study. This chapter proposes new criteria in distinguishing orthologs from paralogs with a phylogenetic approach by using the *NPC1* gene which is responsible for a neurodegenerative genetic disorder: Niemann-Pick type C. Moreover, this chapter describes an algorithm and defines new parameters in missense mutation effect prediction. In Chapter Six, the dissertation is summarized and applications of this work and future aims are discussed.

1.4 References

Aravind L. 2000. Guilt by association: contextual information in genome analysis. *Genome Research* **10**: 1074-1077.

Biro J. 2004. Seven fundamental, unsolved questions in molecular biology: Cooperative storage and bi-directional transfer of biological information by nucleic acids and proteins: an alternative to "central dogma". *Medical hypotheses* **63**: 951-962.

Crick F. 1970. Central dogma of molecular biology. *Nature* **227**: 561-563.

Crick FH. 1958. On protein synthesis. In *Symposia of the Society for Experimental Biology*, Vol 12, p. 138.

Cristianini N, Hahn MW. 2006. *Introduction to computational genomics: a case studies approach*. Cambridge University Press.

De Gobbi M, Viprakasit V, Hughes JR, Fisher C, Buckle VJ, Ayyub H, Gibbons RJ, Vernimmen D, Yoshinaga Y, De Jong P. 2006. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* **312**: 1215-1217.

Donohue J, Trueblood KN. 1960. Base pairing in DNA. *Journal of molecular biology* **2**: 363-371.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**: 496-512.

Goffeau A, Barrell B, Bussey H, Davis R, Dujon B, Feldmann H, Galibert F, Hoheisel J, Jacq C, Johnston M. 1996. Life with 6000 genes. *Science* **274**: 546-567.

Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill DP, Kania R, Schaeffer M, St Pierre S et al. 2008. Big data: The future of biocuration. *Nature* **455**: 47-50.

Kumar S, Dudley J. 2007. Bioinformatics software for biologists in the genomics era. *Bioinformatics* **23**: 1713-1717.

Luscombe NM, Greenbaum D, Gerstein M. 2001. What is bioinformatics? A proposed definition and overview of the field. *Methods Inf Med* **40**: 346-358.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151-1155.

Mount DW. 2001. *Bioinformatics: sequence and genome analysis*. Cold spring harbor laboratory press New York:.

Ohno S. 2013. *Evolution by gene duplication*. Springer Science & Business Media.

Reha-Krantz LJ. 2010. DNA polymerase proofreading: Multiple roles maintain genome stability. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **1804**: 1049-1063.

Satake W, Nakabayashi Y, Mizuta I, Hirota Y, Ito C, Kubo M, Kawaguchi T, Tsunoda T, Watanabe M, Takeda A. 2009. Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. *Nature genetics* **41**: 1303-1307.

Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE. 2015. Big Data: Astronomical or Genomical? *PLoS Biol* **13**: e1002195.

Taylor JS, Raes J. 2004. Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet* **38**: 615-643.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA. 2001. The sequence of the human genome. *science* **291**: 1304-1351.

Watson JD, Crick FH. 1953. Molecular structure of nucleic acids. *Nature* **171**: 737-738.

CHAPTER 2. Literature Review

This chapter summarizes the current literature on technical concepts, which are crucial to prepare the ground for the following chapters.

2.1 Sequence Databases

As discussed in the previous chapter, the first aim of bioinformatics is to store and organize the genomic data. Accumulating sequence data should be publicly available and easily accessible to researchers in order to increase the rate of biological discoveries. There were parallel attempts in storing and presenting annotated nucleic acid sequence data. Three major resources for nucleotide sequence data are:

i. ENA (European Nucleotide Archive) maintained by EBI (European Bioinformatics Institute)

ii. GenBank maintained by NCBI (National Center for Biotechnology Information)

iii. DDBJ (DNA Data Bank of Japan) maintained by NIG (National Institute of Genetics)

These three databases are in collaboration (International Nucleotide Sequence Database Collaboration) and share a spectrum of raw nucleotide data (Brunak et al. 2002). Therefore, in principle, nucleic acids records submitted to any of the databases above should be available through any one of three databases.

The same approach is applied for protein sequences as well. Although many proteins have been experimentally characterized, hundreds of them will never be characterized in laboratories. Therefore, accurately annotated and well-organized protein sequence databases are important for biologists. There are two centers leading and maintaining protein sequence databases: EMBL (European Molecular Biology Laboratory) and NCBI. EMBL offers a comprehensive database, UniProt, which contains variety of protein information besides sequences (UniProt 2015). NCBI offers several databases, each of which is independently served. The following list explains the currently active databases that store, annotate, organize and serve protein sequences.

i. UniProtKB/Swiss-Prot (by EMBL) hosts proteins which were curated and reviewed manually.

ii. UniProtKB/TrEMBL (by EMBL) contain proteins that are automatically annotated.

iii. UniParc (UniProt Archive by EMBL) is a non-redundant database and each protein sequence (collected from different resources) is assigned to an ID.

iv. UniRef (UniProt Reference Clusters by EMBL) contains sets of UniProtKB (including isoforms) and selected entries from UniParc. The sets are non-redundant as homologous proteins are clustered together. The sets of UniRef100, UniRef90 and UniRef50 were defined based on the sequence identity (100%, 90% and 50% respectively) that is used to cluster similar sequences.

v. GenPept (by NCBI) is collection of protein records which are automatically translated from coding sequences in the GenBank resource.

vi. RefSeq (by NCBI) database contains non-redundant and curated set of proteins (Pruitt et al. 2012). A protein record is not repeated in this database for the same genome. It does include alternative gene products, isoforms.

vii. NCBI non-redundant (NR) database is the collection of protein sequences in which identical sequences are clustered together. Though the main resource is GenPept, it is not limited to it. The sequences are also collected from other resources such as RefSeq, Protein Data Bank (PDB), UniProtKB/Swiss-Prot etc.

By definitions UniProtKB/TrEMBL and GenPept should be similar type of databases as they automatically annotate proteins from coding sequences. UniProtKB/Swiss-Prot and RefSeq databases are subsets of UniProtKB/TrEMBL and GenPept respectively. By the same logic, they can be considered equivalent, as their objective is to characterize proteins manually. UniParc and NCBI nr databases can be considered similar because they are collections of unique protein sequences. These two streams of protein sequence resources (maintained by EMBL-EBI and NCBI) share the majority of the sequences, however the content of the databases are not identical. Therefore, the results using the “equivalent” databases may not be the same.

2.2 Comparative sequence analysis

Duplication followed by polymorphisms and selection pressures are the main processes driving the evolution of genes - *de novo* inventions are more infrequent than alterations of existing coding regions (Hughes 1994). Because genes were derived from each other, homologous genes/proteins have similarities in terms of sequence. Therefore, a protein with unknown function can be compared to the existing ones in order to predict potential function, interaction partners and cellular localization. Comparative analysis of DNA and protein sequences has been a successful branch of computational biology since the beginning of the genomics era.

Nucleotide and amino acid sequences are composed of alphabets with the length of four and twenty respectively. One can expect that sequence comparison should not be different than text string comparison from the computational point of view. Although the string and sequence comparison algorithms are comparable, the parameters in molecular biology are different. First, there are indels in evolution, which can be short piece of sequence as well as long stretches. Second, the single point mutations should be taken into account carefully by considering about the likelihood of substitutions. Not all of the amino acid substitutions are equally tolerable. For this reason, there is always a guide in sequence comparison, called substitution matrix. These matrices include scores for each possible amino acid substitutions. The amino acids sharing physiochemical properties are more likely to replace each other. In benchmark alignments of homologous sequences, the statistics of each substitution establish the tendency patterns of amino acid replacements. Therefore, by guidance of the matrix, this kind of substitutions is favored.

Sequences must be aligned in order to match the residues sharing common position in the ancestral sequence. Due to indels, sequences may not be aligned perfectly. Thus, indels are represented by gaps introduced in the alignment. In the alignment algorithms, a gap is introduced with a penalty cost to prevent false/artificial indels. There are two types of sequence alignment serving to different purposes: *pairwise* and *multiple*.

2.2.1 Pairwise alignment

Basically, the aim of pairwise alignment is to answer the question of “Are two proteins homologous?”. Although similarity is a good indicator of homology, it doesn’t necessarily prove it because similarity between two sequences may arise by random chance. Also, sequences may share only partial similarity due to gene duplication, fusion and deletion events. For this reason, there are methods developed to detect positional and general similarities. Thus, there are two types of pairwise alignment: (i) local and (ii) global alignments **(Figure 2-1)**. In global alignment the aim is to align entire length of sequences. On the other hand, local alignment focuses on subsequences that are most similar between two sequences. The aim of local alignment is to find the subsequences, which give the highest score (calculated from the substitution matrix) when aligned. Because introduced gaps cause penalties in alignment scoring, the number of gaps in local alignment is kept at minimum.

```
        Q   K   E   S   G   P   S   S   S   Y   C
A       |       |   |   |                       |
    V   Q   Q   E   S   G   L   V   R   T   T   C
```

```
                E   S   G
B               |   |   |
                E   S   G
```

Figure 2-1 Pairwise alignment types.

A) Global and B) local alignments.

Pairwise comparison of *n* number of sequences with each other is computationally easier than building multiple sequence alignment of *n* sequences. Thus, pairwise alignment calculations are efficient and relatively fast in comparison to multiple sequence alignments. There are tree methods in this type of alignment: (i) Dot-matrix method; (ii) Dynamic programming and (iii) Word method (Rekepalli 2007). Dot-matrix method is efficient to reveal insertions, deletions and repeats. However, it is slow when analyzing large sequences. Dynamic programming, itself, cannot identify insertions and deletions efficiently. “Word” (or k-tuple) remains as the most efficient, and thus preferred pairwise alignment method. By this approach, unnecessary calculations between irrelevant sequences are avoided. Although this method doesn’t result in the optimum alignment, it is fast and thus appropriate to search similar sequences in large databases.

2.2.1.1 BLAST: The popular pairwise sequence comparator

There have been several algorithms developed to perform pairwise alignments in order to calculate similarity score between pairs of genes/proteins, but most of them did not scale efficiently with the number of sequences to be searched. After BLAST algorithm was developed, most probably because of its speed, it became the most popular tool to retrieve similar sequences. It uses the Word method to align two sequences. BLAST algorithm has several subprograms to search different input types against different types of databases **(Table 2-1)**. Because this dissertation focuses on proteins, this chapter specializes in blastp (protein-protein blast) only.

Table 2-1 BLAST programs.

Query	**Subject**	**Program to Use**
Nucleotide	Nucleotide	blastn or tblastx
Nucleotide	Protein	blastx
Protein	Nucleotide	tblastn
Protein	Protein	blastp

The alignment scores are calculated through a substitution matrix of choice such as BLOSUM (Blocks Substitution Matrix) or PAM (Percent Accepted Mutation). PAM is derived from evolutionary distances of proteins. The following number (i.e. 30 in PAM30 matrix) indicates the allowed percentage of substitutions in 100 amino acid-length sequence. PAM250 matrix accepts multiple substitutions per site. Unlike PAM, BLOSUM is not designed based on evolutionary distances. Ungapped alignments of protein families (or domains) are considered as reference. The number followed by BLOSUM (such as 62 for BLOSUM62 matrix) is the indicator of identity threshold when clustering sequence to build blocks.

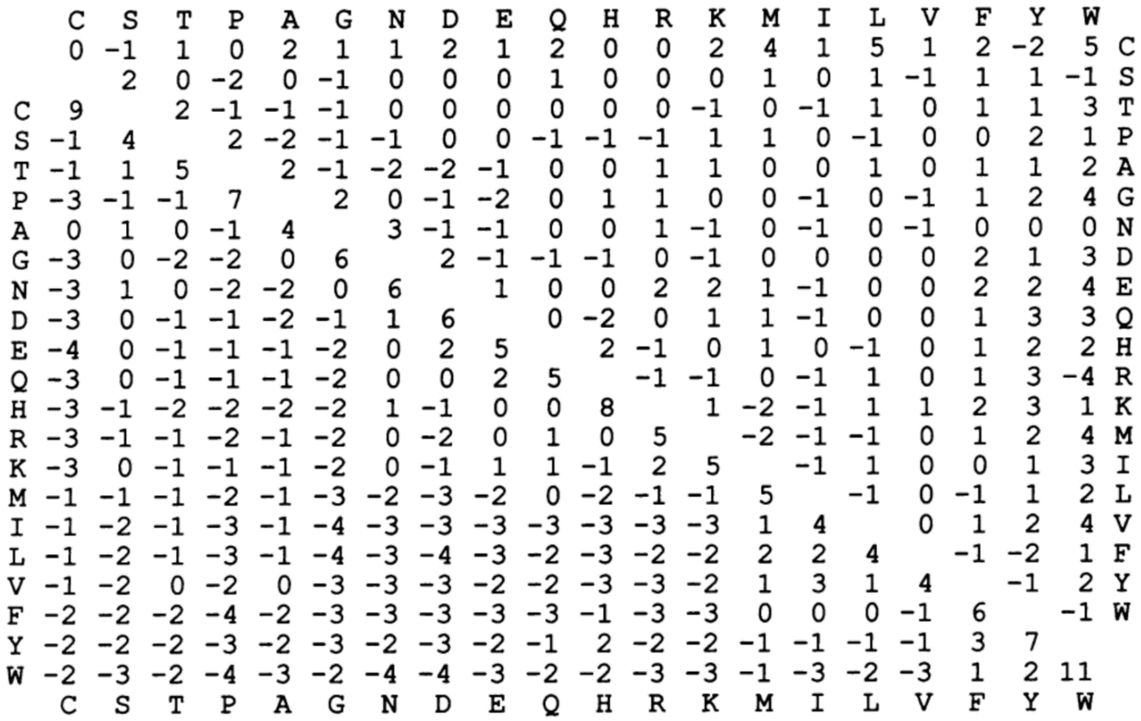
A single matrix cannot be efficiently applied for every case as each BLAST query has its own evolutionary dynamics (Altschul 1991; Altschul 1993). By default, BLOSUM62 **(Figure 2-2)** is utilized by BLAST, which has shown to be most efficient scoring matrix in identifying low similarities in general (Henikoff and Henikoff 1992). For short protein sequences, the scoring matrix should be more conservative for substitutions in order to eliminate false positives. Compared to BLOSUM, PAM matrices have higher mismatch penalties for amino acid substitutions, consequently they are more appropriate for short (less than 50 amino acids) length of queried sequences **(Figure 2-2)**.

NCBI is the primary host for the BLAST algorithm. It enables users to “blast” their sequence of interest on a database of choice (such as RefSeq, NR, Swiss-prot etc.) and against a specified taxonomic level. Therefore, the database to search similar sequences can be manually set. Statistical significance of the hits (found similar sequences matched to query) is indicated by an important parameter called *E-value*. E-value depends on the database size and gives the number of hits to expect due to random chance only (Kerfeld and Scott 2011). Lower E-values means that the match is less likely to be false a positive. For instance, with E-value of 1 indicates one false positive “similar” sequence expected to be found by random chance in the given database.

BLAST algorithm is heuristic and it aims to find similar sequences quickly without the concern of finding optimum alignments. Thus in principle, when two sequences are compared by BLAST, presence of another local alignment with higher score than the match is possible. Also, because it is a local alignment tool, it is not designed to provide high quality global alignments. For these reasons, though it is one of the fastest tools to find similar sequences on large databases, BLAST shouldn’t be used to infer distances between two proteins and it cannot replace the function of multiple sequence alignment tools.

2.2.1.2 PSI-BLAST: To detect distant homology

The problem of pairwise alignment is the fact that only obvious similarities can be detected. Subtle similarities may not be identified through pairwise comparison **(Figure 2-3)**. Two homologous sequences derived from a common ancestor could have diverged enough to loose amino acid identity. However, the physical-chemical composition of sequence may have been conserved to perform the same function, which is apparent in



	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
	0	−1	1	0	2	1	1	2	1	2	0	0	2	4	1	5	1	2	−2	5	C
		2	0	−2	0	−1	0	0	0	1	0	0	0	1	0	1	−1	1	1	−1	S
C	9		2	−1	−1	−1	0	0	0	0	0	0	−1	0	−1	1	0	1	1	3	T
S	−1	4		2	−2	−1	−1	0	0	−1	−1	−1	1	1	0	−1	0	0	2	1	P
T	−1	1	5		2	−1	−2	−2	−1	0	0	1	1	0	0	1	0	1	1	2	A
P	−3	−1	−1	7		2	0	−1	−2	0	1	1	0	0	−1	0	−1	1	2	4	G
A	0	1	0	−1	4		3	−1	−1	0	0	1	−1	0	−1	0	−1	0	0	0	N
G	−3	0	−2	−2	0	6		2	−1	−1	−1	0	−1	0	0	0	0	2	1	3	D
N	−3	1	0	−2	−2	0	6		1	0	0	2	2	1	−1	0	0	2	2	4	E
D	−3	0	−1	−1	−2	−1	1	6		0	−2	0	1	1	−1	0	0	1	3	3	Q
E	−4	0	−1	−1	−1	−2	0	2	5		2	−1	0	1	0	−1	0	1	2	2	H
Q	−3	0	−1	−1	−1	−2	0	0	2	5		−1	−1	0	−1	1	0	1	3	−4	R
H	−3	−1	−2	−2	−2	−2	1	−1	0	0	8		1	−2	−1	1	1	2	3	1	K
R	−3	−1	−1	−2	−1	−2	0	−2	0	1	0	5		−2	−1	−1	0	1	2	4	M
K	−3	0	−1	−1	−1	−2	0	−1	1	1	−1	2	5		−1	1	0	0	1	3	I
M	−1	−1	−1	−2	−1	−3	−2	−3	−2	0	−2	−1	−1	5		−1	0	−1	1	2	L
I	−1	−2	−1	−3	−1	−4	−3	−3	−3	−3	−3	−3	−3	1	4		0	1	2	4	V
L	−1	−2	−1	−3	−1	−4	−3	−4	−3	−2	−3	−2	−2	2	2	4		−1	−2	1	F
V	−1	−2	0	−2	0	−3	−3	−3	−2	−2	−3	−3	−2	1	3	1	4		−1	2	Y
F	−2	−2	−2	−4	−2	−3	−3	−3	−3	−3	−1	−3	−3	0	0	0	−1	6		−1	W
Y	−2	−2	−2	−3	−2	−3	−2	−3	−2	−1	2	−2	−2	−1	−1	−1	−1	3	7		
W	−2	−3	−2	−4	−3	−2	−4	−4	−3	−2	−2	−3	−3	−1	−3	−2	−3	1	2	11	
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Figure 2-2 BLOSUM62 (below) and PAM160 (top) substitution matrices.

Both matrices have identical entropies. Copied from (Henikoff and Henikoff 1992).

homologous sites with reduced alphabets. A functional unit of a protein have a *profile* which carries the characteristics of the represented unit. Remotely related and functionally equivalent proteins share alphabet characteristics (profile) to perform the same function. Thus, a profile can be used as a signature for detecting remote homologies.

To overcome the sensitivity problem of pairwise comparison, PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool) was developed. The algorithm is processed iteratively. First iteration is identical with the conventional BLAST procedure. A query sequence is blasted against a database. Before next iteration, significant BLAST

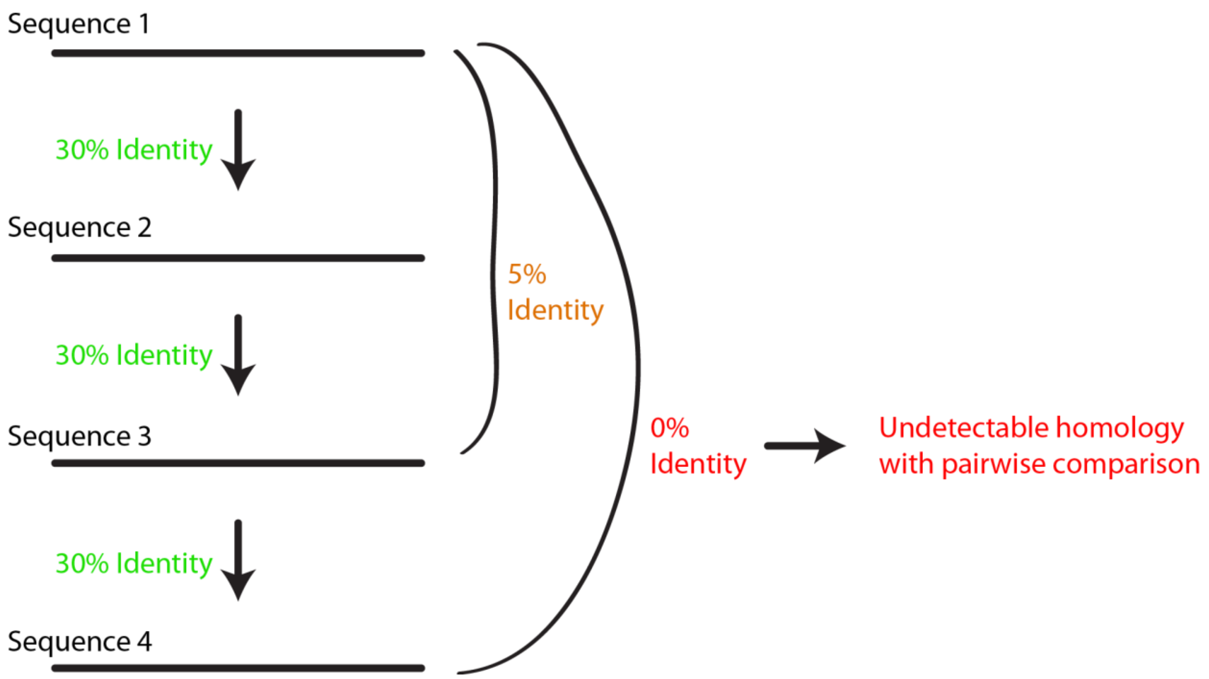


Figure 2-3 The problem of pairwise comparison in detecting remote homology.

Sequence 1 and Sequence 4 are distantly related with no sequence identity. It is not possible to identify the relationship by comparing these two sequences only.

hits (under a specified e-value) are collected and aligned. From the alignment, a *profile* is generated. The profile includes observed amino acid counts for each position in the alignment. These counts are used to determine the likelihood of observing a certain amino acid at a certain position. This profile is called PSSM (Position-specific scoring matrix). PSSMs contain the conservation pattern of the sequence they represent. In the second iteration, the newly generated PSSM is used as query. Compared to pairwise alignment, PSSM is more sensitive to find similar hits as it contains information from not a single sequence but multiple sequences. After collecting the results, new significant hits are added to the alignment, which is used to generate a newer PSSM. For each iteration, PSSM is recalculated with the newly selected sequences and queried against the selected database. After each iteration new hits can be found or new hits may not be found after some point. Thus, the process is iterated until the point of satisfaction with the results, an arbitrary iteration number, or the point that no new sequences can be captured. It is important to note that PSI-BLAST algorithm, as other BLAST tools, is heuristic i.e. it doesn’t guarantee to yield the optimal solutions.

Like other BLAST tools, PSI-BLAST is also available at NCBI webserver. Additionally, the graphical user interface is convenient to apply PSI-BLAST. In every iteration, users are given an option of which hits to include and exclude in the computation of the following PSSM. In PSI-BLAST, hits are not guaranteed to be sorted from most to least similar. For this reason human intervention is important. In this case deciding on what to include is a challenge. Statistical significance is an important criterion to be careful about. When manually performing PSI-BLAST on NCBI server, if a hit is at the border of statistical significance, there are consequences of including and excluding the hit depending on the truth of match. If the hit is false positive, inclusion of an irrelevant sequence may make the next PSSM more diverged, which would result in more false positives at the next round. If the hit is true positive, but a diverged one, excluding that sequence in PSSM may end the iteration at that point and more distantly related proteins may never be found.

2.2.2 Multiple Sequence Alignment

Multiple sequence alignments (MSAs) are built to compare three or more DNA or protein sequences with each other. Usually, sequences that are related to each other with a common ancestor are aligned to reveal common patterns as homologous sites. Moreover, MSAs are prerequisites for most of phylogenetic analyses. The aim of MSA is to align homologous residues, such as amino acids which used to have the same common ancestor. Thus, columns in the alignment should ideally be homologous residues in each position. The length of sequences can be different in homologous sequences due to insertions and deletions. For this reason gaps are introduced and represented with hyphens.

MSAs can be built manually or automatically. Because manually aligning sequences is tedious especially for large data, computational algorithms were developed to perform this task. Automatic MSA building is an extensive research topic in the field of computational biology. Few popular MSA tools are: CLUSTAL W, T-Coffee, MUSCLE and MAFFT.

As homologous sequences are derived from a single ancestor, every residue in a sequence had an ancestor residue (except for recent insertions). Although in theory there should be a single correct MSA having only homologous residues aligned, in practice this is not possible yet. Computational algorithms generate more than one MSAs. The “best” MSA determined by specific criteria is selected to be the optimum one. However, in most cases, the optimum MSA does not identically reflect the true alignment. Thus, the achieved MSA is an approximation and it usually contains mistakes. Manual refinements should be applied on automatically generated MSA.

Substitution matrix is the most important criterion in building MSAs. In protein sequence alignment, the substitution matrices are based on the observation of current substitutions. Usually, amino acids sharing one or more physical or chemical property are more likely to substitute each other. For instance, S and T are often replaceable because they are small and polar. Additionally, D and E (negatively charged and acidic); H, K and R (positively charged); F, I, L, M, V (hydrophobic); F, Y (aromatic) are often interchangeable when only their physiochemical characteristic is conserved at a certain position. Cysteine

(C) on the other hand is not replaceable with any amino acid when it builds a disulfide bond which is crucial for the protein structure. In DNA alignment, application of a substitution matrix is not possible because of the small size of the nucleotide alphabet. For nucleotide alignment there is ~1/4 random chance of having an identical match. Therefore, in DNA alignment only identity is used as the substitution matrix: Adenine (A) can be favorably matched only with A. For this reason, nucleotide alignments are often ambiguous and result in more mistakes compared to protein sequence alignments. Therefore, if possible, instead of aligning DNA of a coding region, building MSA on amino acid sequence yields more reliable results. If nucleotide sequence of a coding region is needed for the analysis, building DNA alignment generated from protein alignment (reverse translation) can be used.

2.3 Phylogenetics and Taxonomy

In biology, it is important to know about the historical relationships between species or inherited biomolecules. This can be achieved by DNA and/or protein comparison or manually curated classification. This chapter introduces basic approaches of hierarchical classification of organisms and biological sequences.

2.3.1 Phylogenetics

Phylogenetics is the study of the evolutionary history of organisms (or their hereditary components) achieved via genetic material. Typically, a phylogenetic tree is inferred from a sequence alignment. Alignments ideally contain only homologous sequences and residues in an alignment column share a common ancestor. A phylogenetic tree should represent the evolutionary history of subjects while explaining the alignment it is derived from. There is a single correct tree, but it is not easy to achieve. Calculations of the likelihoods of all possible tree topologies are computationally impossible with the current hardware for more than tens of sequences. So, the aim of algorithms is to pick the best tree among calculated ones.

2.3.1.1 Methods

There are four major methods to build phylogenetic trees: (i) distance matrix; (ii) maximum parsimony; (iii) maximum likelihood and (iv) Bayesian.

The distance method depends on relative evolutionary distances between each pair of sequences to be compared. The distance can be achieved through pairwise comparison by which a genetic distance matrix is established (Felsenstein 1988). Pairwise comparison results in an underestimated genetic distance as they count each polymorphism as a single mutational event, which may actually be caused by multiple events (Salemi and Vandamme 2003). Underestimated dissimilarity score can be converted to an evolutionary distance by an adjustment formula (Jukes and Cantor 1969). Then, tree topology is inferred from estimated distances. There are a few types of tree-building approaches based on evolutionary distances such as the clustering approach and the neighbor-joining method.

UPGMA (Unweighted Pair Group Method with Arithmetic Mean) is one of the clustering methods that provides a hierarchically clustered tree. This algorithm assumes that all sequences to be compared are at an equal distance from the root. Therefore, all tips of the individual branches are aligned **(Figure 2-4).** In reality, this is hardly the case. Two biological species are rarely at the identical distance from their closest common ancestor, as their evolutionary rates are likely to be different. For this reason UPGMA is rarely used to understand true evolutionary relationships.

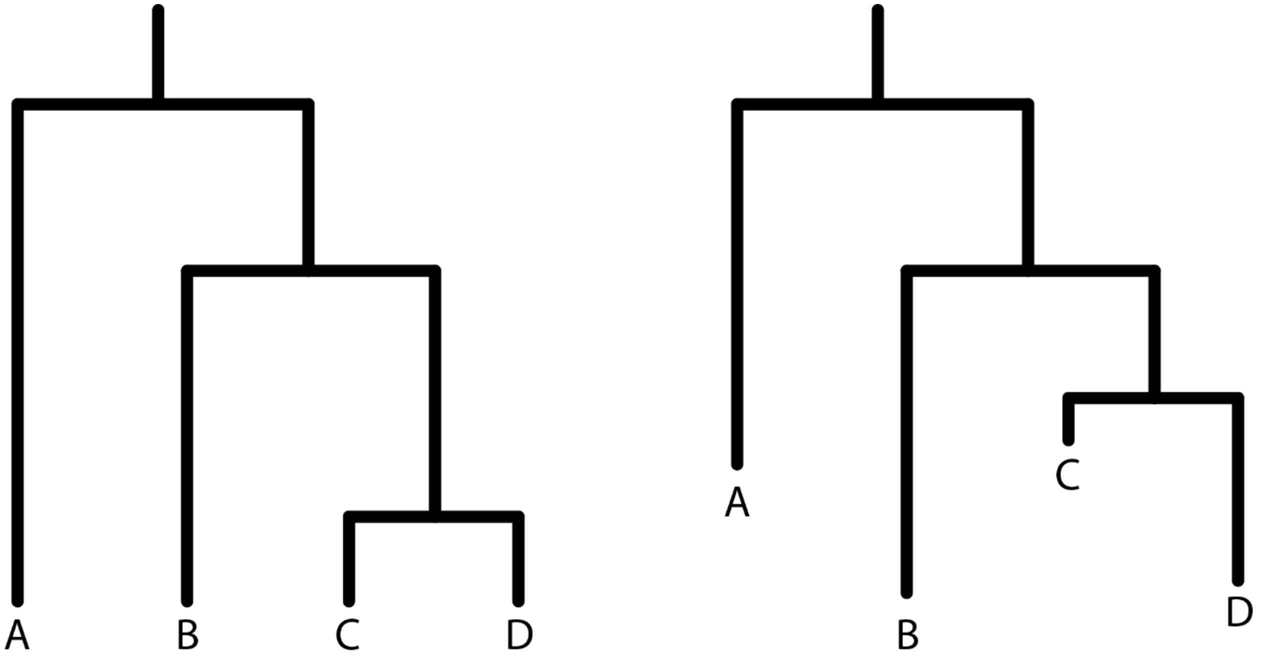


Figure 2-4 Cladogram vs phylogram.

Left panel shows a cladogram which UPGMA results in. Right panel shows a phylogram.

Neighbour-joining (NJ) is another distance-based method that is computationally efficient. The algorithm clusters the two closest taxa and considers the cluster as a single node to compare with others and additively constructs the tree. In order to achieve the correct phylogenetic tree with NJ approach, distance matrix must be statistically consistent. In other words, the additive algorithm should progress with matching distances in the matrix. Because distance matrix generating algorithms are approximations, the true distance table cannot be produced easily. Thus, the NJ approach may also fail to yield the correct tree. It is preferable over more accurate non-distance-based tree inferring methods due to its efficacy in computing. A tree from thousands of sequences can be built with NJ in a short time on a single node. It is also preferable over the UPGMA approach because it doesn't assume that all sequences have evolved at the same rate, thus it results in non-equal branch lengths from the root.

Maximum parsimony approaches assume a minimum evolutionary event number to achieve the observed data. It assigns common ancestors to each node in a way that it favors the minimum substitution number. Although the logic behind the method is reasonable for morphological properties, for DNA/protein sequences, it is not well suited. In sequence alignments there is more than one way to achieve minimum evolution, which may result in multiple optimal trees. With morphological data, which contains more complex structures that are less likely to evolve independently compared to sequence residues, maximum parsimony performs well.

Maximum likelihood (ML) is a statistically well-understood method to estimate parameters of an evolutionary model that explains the evolutionary process in which the observed data went through. In phylogenetics, this approach can be used to evaluate the pattern of branching by considering probabilities explaining the observation under a given model. ML is a powerful tree-inferring method as it provides biologically meaningful trees compared to parsimony and distance-based approaches. However, it is computationally intense. Because searching every possible tree may not be plausible, a heuristic method is usually applied. Even the heuristic approach doesn't enable applying ML on large data sets containing over thousand sequences due to computational constraints, unless there is a chance of using a supercomputer.

Bayesian inference is another method to generate trees from sequence alignments. It is similar to ML because it also uses a likelihood function and results in the probability distribution of possible trees. However, the difference of this method is that it uses prior knowledge on the data (prior probability) to calculate posterior probability. As a heuristic method it doesn't guarantee the best tree. However, it is used as widely as the ML method. One main disadvantage of this method is the lack of speed given the nature of the algorithm. For small sets of sequences Bayesian inference works well.

2.3.1.2 Which One to Use: Protein or DNA Sequence?

DNA is the source of information that is needed to synthesize proteins, whereas proteins are the fundamental functional elements in molecular biochemistry, which are subject to natural selection rather than protein coding DNA regions. Coding DNA can be modified as long as protein function and expression are maintained. For reasons explained below, using protein sequences in phylogenetic analysis is more appropriate to obtain trees closer to the "true" tree.

1) *Multiple codons one amino-acid*

As discussed in the introduction same amino-acid can be encoded by multiple codons. For this reason, any codon producing the required amino acid will not be under an eliminating selection pressure. So, different DNA sequences may result in the same fitness.

2) Codon usage bias

Organisms may have different preferences for certain codons which encode for the same amino acid. Thus, comparing DNA sequences from different organisms may reflect an artificial change in nucleotides which have no effect at the protein level.

3) "Almost" universal genetic code

Genetic code is general and applies to nearly all organisms with exceptions. Unusual genetic codes were reported in lower eukaryotes (Horowitz and Gorovsky 1985; Salemi and Vandamme 2003). Non-identical (or even non-equivalent) codons may result in same amino acid sequences.

4) Higher noise in DNA

In a DNA sequence there are only four letters and at any position of the alignment there is 25% chance of matching nucleotides randomly. Therefore, the probability aligning non-homologous residues is much higher in DNA compared to proteins which would have ~5% probability of "by-chance" misalignment. Therefore, by using proteins, it becomes more likely to align homologous residues (the amino acids that share a common ancestor at the same position).

5) Non-coding regions within a gene

Genes include regions called regulatory sequences which are not transcribed. In eukaryotes, there are introns which are removed at pre-mRNA level after transcription. In addition, there are untranslated mRNA regions. Therefore, in order to infer meaningful phylogenetic trees these regions should be removed from the DNA alignment, as they can be highly divergent and cause substantial unnecessary noise in the MSA. A high-quality gene alignment at DNA level can be performed using only coding regions.

On the other hand, there are minor drawbacks of using protein sequence. First, positive/negative selection cannot be quantitatively identified using only protein sequences. Second, non-coding sequences may contain conserved information and they are only accessible at the DNA level. And finally, in eukaryotes, alternatively spliced proteins result in variations of in the products of the same gene. Thus, even if the protein sequences are originated from the same gene, they cannot be aligned well at splice site regions. The last drawback is addressed in chapter 5.

2.3.1.3 Reading phylogenetic trees

A connected sets of organisms in trees is called a *taxon* (plr. *taxa*). Each tip (often called leaf) represents an organism (or gene/protein) while each node represents the common ancestor of the descendants **(Figure 2-5)**. Phylogenetic trees reflect not only clusters of similar taxa but also evolutionary relationship between them, which is the point of difference between phylograms and cladograms. Evolutionary trees show how ancestors are related with their descendants quantitatively. In a vertically aligned tree, the distance between two taxa is measured by the total vertical branch length **(Figure 2-5)**. A clade is a set of organisms that share an ancestor whose all descendants are in the set **(Figure 2-5)**. Clades are composed of minimum two organisms. Small clades can be nested in large ones.

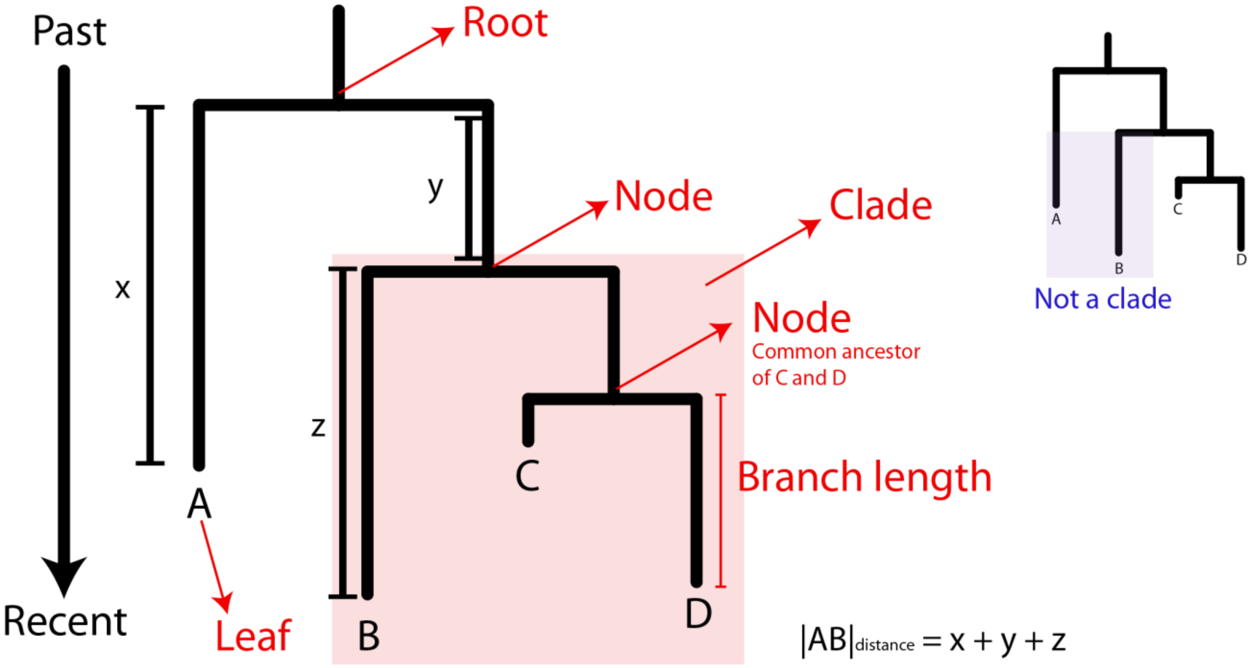


Figure 2-5 Reading phylogenetic trees: illustrated terms and measuring distances between two leaves.

2.3.2 Taxonomy

Taxonomic trees reflect hierarchically categorized organisms and aim to represent evolutionary classification. There are several different attempts on building taxonomic classification. Most of these databases are specialized in a certain clade of organisms such as bacteria, plants etc. NCBI offers the most comprehensive taxonomy database with the attempt of including all organisms from each domain of life (including viruses and viroids). This database is carefully built and maintained by considering not only DNA/protein similarities between organisms, but also consensus literature in the classification. Ranking categories start from root (which includes every organism with no exception) to subspecies and every kind of other rankings between them. Each species and classification node is given a “taxid” which is a unique identifier. Every node including species can be traced back to the root with the lineage information. One drawback of the classification is that ranking levels from root to species are not standard. One organism may have only 8 levels described, whereas others may have 16. So, not every organism has the same hierarchical levels.

Resolving ancestral relationships in early eukaryotic domain of life remains a challenge (Burki et al. 2007). There is no consensus of how deep eukaryotic groups are connected to each other (Hampl et al. 2009). Although their evolutionary relationships couldn’t be established, there are 5-6 eukaryotic supergroups known to be diverged at early stage of eukaryotic evolution. These supergroups are: Unikonta (Opisthokonts and Amoebozoa), Chromalveolata, Plantae, Excavata and Rhizaria (Koonin 2010). Because their hierarchical relationships are arguable, eukaryotic supergroups are not used in taxonomic database. However, supergroup assignment to genomes would be useful to detect the genes that are common in all of them, which is an indicator of the presence of the genes in the last eukaryotic common ancestor (Koonin 2010).

Generally, taxonomic databases provide a good overview of which organisms are similar to each other, however it has low resolution within clades. For instance, a question of which species within a genus are more similar to each other cannot be answered unless they are classified with an intermediate rank between genus and species.

2.4 Protein Domains

Amino acids are the building blocks of proteins. Linked amino acids (polypeptide chains) fold and give rise to three dimensional protein structures. Proteins are composed of one or more functional and structural units, called *domains,* which evolve and function independently from the rest of the protein. Domain sizes are limited. With 100 amino acids in average, majority of domains are shorter than 200 and longer than 40 amino acids (Islam et al. 1995; Jones et al. 1998; Wheelan et al. 2000).

As structural units, domains are not fully conserved sequence-wise. A *motif* is found in a domain, a set of multiple amino-acid residues, which is probably the most critical sequence pattern for the biological function of the protein. Because they are actively involved in the function (such as catalytic activity, binding, protein-protein interaction etc.) motifs are highly conserved and cannot be mutated without a cost in fitness. Residues of motifs are structurally in close proximity with each other and most of them are closer to each other at the sequence level. Unlike domains, motifs cannot be classified as structural units. A *fold* on the other hand, represents a domain or a smaller unit of a domain from the structural point of view only. Folds are composed of multiple secondary structural elements. A domain is composed of a single or more folds.

Single domain proteins (SDPs) are more often found in the early stages of life where MDPs are considered as more recent as they are derived from SDPs (Di Roberto and Peisajovich 2014). Evolution of new protein functions occurs through combining and rearranging domains. Organismal complexity positively correlates with the abundance of multi-domain proteins (MDPs) (Di Roberto and Peisajovich 2014). MDPs are produced by evolutionary mechanisms such as duplication, fusion, disassociation and divergence (Kummerfeld and Teichmann 2005). Though it is possible to invent a domain from structurally disordered sequences, this mechanism remains minority among others (Moore and Bornberg-Bauer 2012; Di Roberto and Peisajovich 2014).

In functional networks, domains play important roles because of their characteristics. Firstly, a domain is an independently-evolvable unit, thus it can exist in various domain organizations. In different compositions, sub-functionalization may occur which results in function divergence of the entire protein. Secondly, many domain functions are extensively regulated. Besides getting turned on/off, the activities vary based on molecule-binding and post-translational modifications. Due to their flexibility in adapting to proper function, domains are highly dynamic in terms of evolution. Useful domains adopted through either vertical or horizontal evolution are kept because of advantage they contributed in the fitness of organisms.

Domain fusion is one of the major mechanisms in generating a new function. Interacting domains on different poly-peptide chains can be fused to act together. Fusion saves cost in protein synthesis and molecular transport, and thus fusion of interacting SDPs is favored in the evolution. Therefore, domains in MDP, if they exist as single proteins, it is very likely that they do interact with each other. This is one way to deduce protein-protein interactions *in silico*. However, not all domains are evenly involved in various domain combinations, and some of them even don’t easily fuse with another one as they are highly conservative in domain architecture preference. These domains are more conserved than the ones freely involved in various domain architectures.

To conclude, protein function evolution and new protein invention heavily depend on the dynamic nature of domain evolution. For this simple reason, understanding domain evolution contributes to understanding protein evolution and thus function.

2.4.1 Domain search methods

Pairwise sequence comparison is not sensitive in identifying homology **(Figure 2-3)**. Homologous sequences share common characteristics, such as conserved residues, motifs, secondary structure patterns etc. These characteristics are reflected on the sequences and easily detectable through MSAs. MSA can be summarized by *profiles* that represent a family of related (homologous) sequences by containing their conserved, and thus important characteristics. Profiles are used frequently to detect homology more sensitively.

One type of profiles is amino-acid frequency matrix, PSSM, which is used in PSI-BLAST algorithm. A domain alignment is converted into a PSSM which can be queried to search for new sequences belonging to the domain family represented by the given matrix. A sequence can also be queried against PSSMs to scan which domains are present in the protein. RPS-BLAST (Reverse PSI-BLAST) is a tool enabling search of sequences against a PSSM database. Matrices contain frequencies of amino acids in each position of proteins. The complete information of an MSA cannot be represented by a score matrix. PSSM profiles are strict about insertion and deletions which reduces applicability for length wise diverging domains. Because of these limitation in substitution matrix-based profiles, probabilistic profiles were developed.

The most popular of these profiles uses a statistical method called Hidden Markov Model (HMM). HMMs take into account the likelihood of every possible transition state categorized as *match, insertion* and *deletion.* Thus, HMMs contain information of the probabilities of state changes as well as the frequencies of each residue in an alignment column. Each position in an HMM has two types of probability: *transition* and *emission*. Transition probability is simply the likelihood of transitions between ‘match’, ‘insertion’ and ‘deletion’ states. Emission probability is the likelihood of observing an amino acid or an insertion at each position. HMMs are more complicated than PSSMs, as they store more information and offer more sensitive domain identification. Sequence of interest can be compared with HMM of a pre-compiled domain/protein family in order to question the relationship. This is a widely used approach to identify protein domains. HMMER3 is a well-known package of HMM profile-related tools (Mistry et al. 2013). HMMscan tool allows searching sequences against a profile database whereas HMMsearch is used to search a certain profile against a sequence database. HMMbuild generates HMM out of a given MSA.

HMMs substantially contributed in protein/domain classification and identification. However, in the investigation of distant homology, the sensitivity of the HMM-sequence comparison still remains under the desired level. A more sensitive approach was developed, which is HMM-HMM comparison. First, a sequence of interest is searched in a sequence database to find closely related sequences using pairwise similarity searches. Significant hits are collected and aligned. The built MSA is then converted to an HMM profile. Next the HMM profile is queried against the same sequence database. Because an HMM is used now, the search is likely to result in newer hits. The new significant hits are also added to the alignment, which is then converted to a second HMM and the process is repeated several times **(Figure 2-6)**. The final HMM profile is then compared with a collection of HMMs. The HMM-HMM comparison **(Figure 2-7)** helps in detection of distant homologies. HHsearch is a popular tool that is often used for HMM-HMM comparisons.

2.4.2 Protein/domain profile databases

Domain profiles are annotated and organized in different forms in several databases. The most popular domain profile database is Pfam (Protein Families) (Finn et al. 2014). Pfam contains MSAs and derived HMM profiles for domains. HMMs in Pfam are built using HMMER3 (Mistry et al. 2013). Pfam-A is the primary collection containing manually annotated protein/domain families. Pfam-B is another resource for other families which are built automatically from protein/domain clusters retrieved from ADDA (Automatic Domain Decomposition Algorithm) (Heger et al. 2005) database. These protein clusters are not manually checked and thus not annotated. Pfam-B domains are generated from protein/domain clusters which don’t overlap with any current Pfam-A models. Therefore, the aim of Pfam-B is to cover “orphan” sequences, however they don’t give any biological insight other than the conservation characteristic of the region.

TIGRFAM is another HMM-based database as it also uses HMMER3 to build models (Haft et al. 2013). Unlike Pfam, it is not domain-oriented. TIGRFAM includes profiles of full-length proteins. Therefore, both SDPs and MDPs are represented by single profiles. TIGRFAM is useful to determine protein families but it is not primarily used to assign domains within proteins.

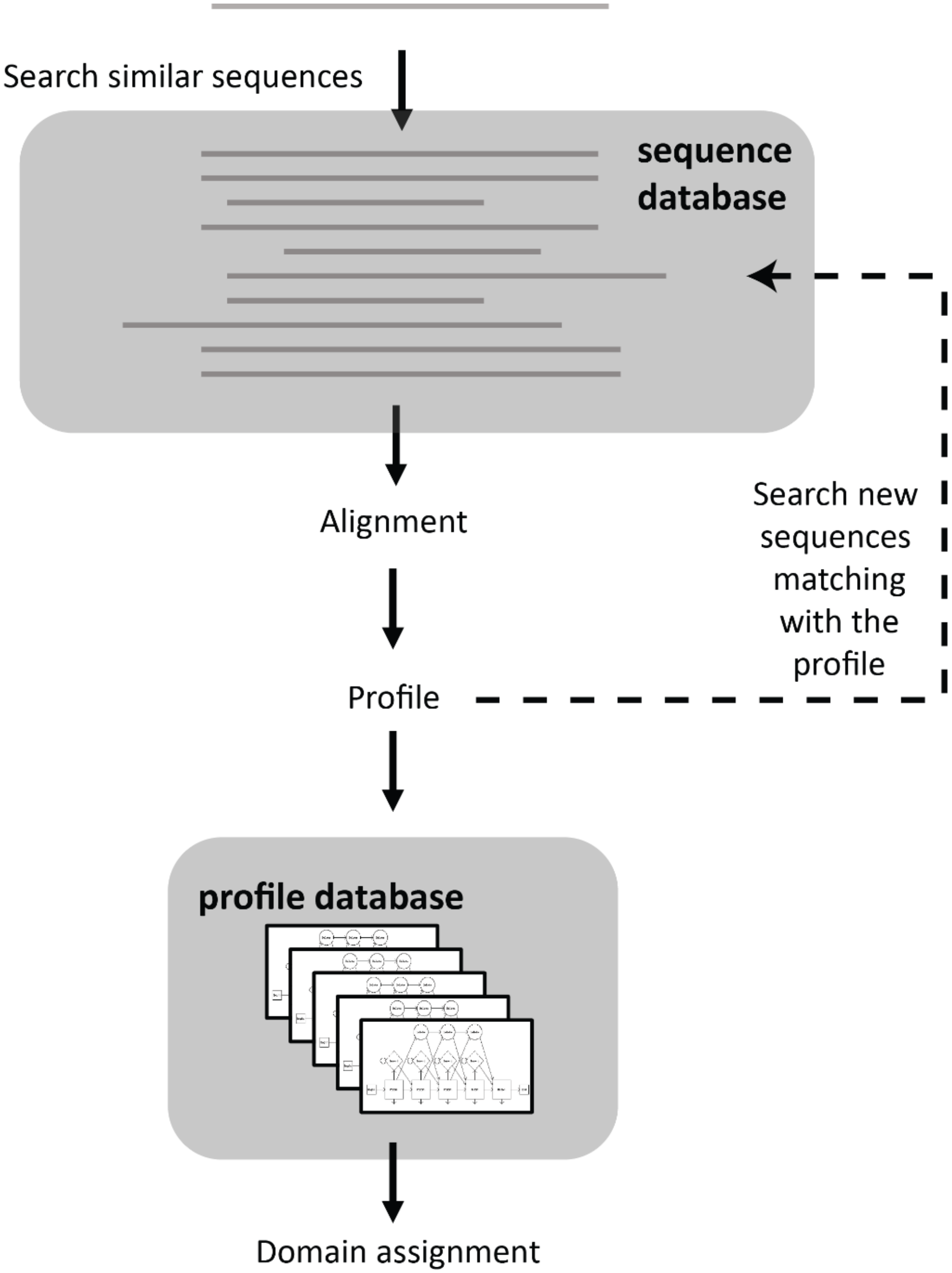


Figure 2-6 Domain assignment through HMM-HMM comparison with a sequence as starting query.

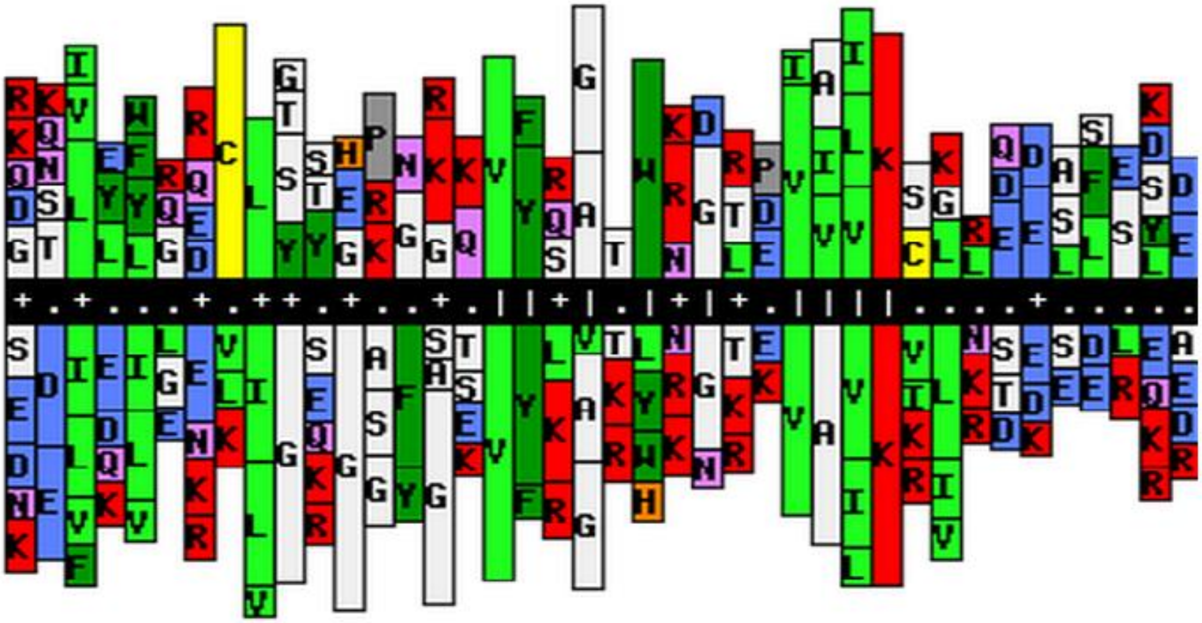


Figure 2-7 A representative scheme of HMM-HMM comparison.

(Copied from HHpred website)

SMART (Simple Modular Architecture Research Tool) database is another domain collection composed of HMM profiles (Letunic et al. 2015). The most important feature of the database, which makes it different than other resources, is that the domains are manually and carefully annotated, and external links to variety of databases are provided. The latest release (version 7) of the database contains only 1204 distinct models.

PIRSF is a database of protein classification that is manually curated (Nikolskaya et al. 2006). Only proteins sharing full-length similarity and having same domain architectures are clustered together. However, as other manually curated databases, it is not comprehensive. Moreover, length-wise diverged protein sequences are not clustered together which causes a drawback of losing homology between them.

COG (Cluster of Orthologous Groups) database is semi-automatically generated and includes clusters of protein and domains which are likely to be orthologous (Galperin et al. 2015). Because the clusters are built automatically, they also include paralogs. The algorithm of COG is based on pairwise sequence comparison in the context of individual genomes. It is different than a regular domain database as only orthologous domains are clustered together. So same domain can be classified in multiple COGs if it is involved in various domain architectures.

Another frequently used database is CDD (Conserved Domain Database) (Marchler-Bauer et al. 2015). Unlike Pfam, TIGRFAM and SMART it is a database of PSSMs. PSSMs are built from manually curated MSAs and sequences are compared with profiles via RPS-BLAST. Domain boundaries are primarily determined by available protein structures, if one exists. CDD also includes PSSM profiles built from MSAs which were retrieved from external databases such as Pfam, SMART, KOG, COG and TIGRFAM.

SCOP2 is an upgraded version of SCOP (Structural Classification of Proteins) database. The database is a collection of hierarchically clustered domains. The classification principle is structure-oriented. The hierarchical classification levels are as the following:

- ➔ **Structural Class:** All alpha
    - ➔ **Fold:** Globin-like
        - ➔ **Superfamiliy:** alpha-helical ferredoxin
            - ➔ **Family:** pyrimidine dehydrogenase N-terminal domain-like
                - ➔ **Protein:** Dihydropyrimidine dehydrogenase, DPYD

SCOP2 also contains crosslinks on relationships between proteins in terms of evolution and structure (Andreeva et al. 2015). The database is manually curated, consequently reliable, however not comprehensive. SUPERFAMILY is an HMM-based database built on the SCOP domains at the superfamily level (Gough and Chothia 2002).

CATH is another structure-oriented collection of hierarchically classified domains (Sillitoe et al. 2015). It is semi-automatically maintained by using entries from PDB (Berman et al. 2000). CATH is an abbreviation representing four levels of classification: (i) **C**lass (equivalent to SCOP Structural class); (ii) **A**rchitecture (equivalent to SCOP fold); (iii) **T**opology; (iv) **H**omologous superfamily (equivalent to SCOP superfamily). The database is built on solved structures. Gene3D is a collection of CATH classifications applied to protein sequences by revealing similarities between solved structures and other sequences whose structures have not been solved (Lees et al. 2014).

SeqDepot is a comprehensive collection of various databases on a non-redundant set of protein sequences and their associated components (Ulrich and Zhulin 2014) . The motivation behind the database is to avoid unnecessary recalculations on predicting protein components for identical sequences. Each sequence in the database is represented by an id (20 characters), which is used to retrieve pre-computed sequence features such as hits from 18 resources including Pfam, TIGRFAM and SUPERFAMILY.

2.4.3 Web-servers for domain searching

Pfam website enables domain searching by HMMER3. It also includes a large repertoire of pre-computed domain architectures on UniprotKB sequences. Pre-computed domain architectures can be queried by entering desired presence and/or absences of domains. Phyletic distribution of single domain presence is also offered. However, genomic absences of domains are not shown.

In Pfam, related domain families are clustered together with profile-profile comparisons such as HHsearch (Soding 2005) and SCOOP (Bateman and Finn 2007). Also the fact that a same region of a sequence has significant hits for more than two Pfam profiles is considered as an indicator of related profiles, thus they are categorized as candidates to be clustered. These clusters are termed *clans* in the database. Though domains which are members of the same clan are highly likely to share a common ancestor, the homology is not guaranteed.

HMMER web server also allows domain prediction through HMMscan (Finn et al. 2015). Unlike Pfam it doesn’t store pre-computed domain architectures. It also allows HMMsearch by querying an MSA. MSA is converted to an HMM automatically with HMMbuild and generated HMM is searched on the database of choice such as Uniprot, SwissProt and PDB.

CD-Search webserver allows querying sequences against CDD (or other compiled PSSM libraries) which is a collection of pre-calculated PSSMs (Marchler-Bauer and Bryant 2004). The search is performed by RPS-BLAST. The web server also enables querying similar domain organizations through CDART (Conserved Domain Architecture Retrieval Tool) (Geer et al. 2002).

SMART is another web resource for searching a sequence of interest on its own database. The most important feature of the SMART website is that it allows querying domain composition and organization (Letunic et al. 2015). By this feature a user is enabled to search for other proteins that have the same composition of the user’s input. Domain organization search allows finding proteins having only the domains in the same order with the input. Taxonomic distribution of the hits (only presence) is also offered.

HHpred is a platform enabling more sensitive domain investigation (Hildebrand et al. 2009). The main goal of the tool is to utilize HMM-HMM comparison to infer remote homologies. The user enters a single input which is iteratively searched against Uniprot or NR database to find similar sequences. This search is performed by either PSI-BLAST or HHblits (Remmert et al. 2012) algorithms, by generating a profile after the first run, followed by the same procedure several times until a pre-determined iteration number is reached. At the end of the iterative sequence search, sequences found to be related are aligned and an HMM is built out of MSA. The generated HMM is queried against a collection of HMMs via HHsearch algorithm. Libraries of HMMs are built from databases retrieved from several resources such as Pfam, PDB, SCOP, CDD, Panther, PIRSF, COG, CATH and SUPERFAMILY. The web server also offers an option to limit the search with one of the several representative genomes. By default, HHpred also uses a contribution of secondary structure information when detecting similarity between profiles. Secondary structure information can be retrieved from either a computational prediction or a structure database (such as PDB).

HHpred web-server allows a single query at a time. If the query is a single sequence (not MSA), similar sequences are searched via HHblits by default. For multi-domain proteins similarity search is biased towards finding the sequences with a similar domain architecture (Hildebrand et al. 2009). Moreover, the profile built at every step of iterative process will represent the domain architecture of the initial query. The biased domain architecture will be specific to find domains which are similar to ones found in the same domain organization. More diverged domain sequences, especially the ones that are found in different domain organizations, are less likely to be identified. Therefore, in this case, the results do not reflect the true sequence diversity of the protein family. This causes a narrower spectrum of the HMM profile and, consequently leads to lower sensitivity than theoretical limit **(Figure 2-8)**. If possible generating profiles out of individual domains would be more sensitive to find more distantly related sequences.

It is worth to note that though there are databases containing pre-calculated HMM-sequence comparison results, there is no such database for HMM-HMM comparisons. Because HMM-HMM comparison is computationally expensive and there is no standard procedure to build first HMM to compare against a collection. With the current hardware and software capabilities, HMM-HMM results can only be achieved through via “on the fly” computations.

2.4.4 Other protein components

Low-complexity regions (LCRs) are subsequences composed of biased amino acid composition with a little degree of diversity (DePristo et al. 2006; Coletta et al. 2010). Most of them are sequences composed of one or few different amino acids, or in other words, it utilizes only a limited subset of the amino acids alphabet. LCRs are usually non-conserved, irregularly spaced and repeated. About 12% of Uniprot amino acids are found to be within LCRs (UniProt 2008). These sequences become non-globular/disordered structures when translated. In sequence similarity searches including profile-sequence comparisons, LCRs cause false positive matches. Therefore, hits on the LCR part of a sequence should be taken into account cautiously. SEG is an algorithm used to predict LCRs (Wootton 1994).

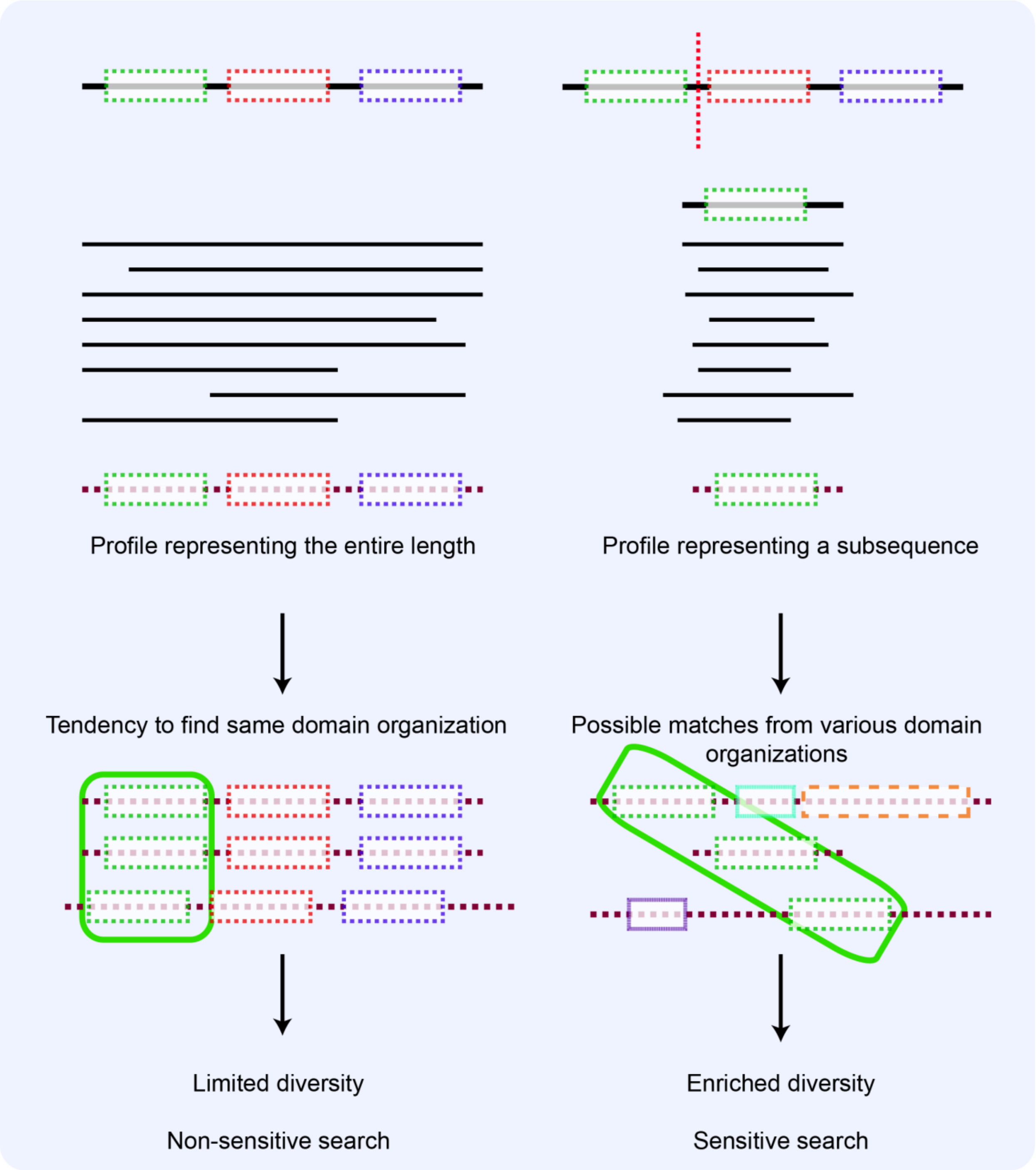

Profile representing the entire length
Profile representing a subsequence
Tendency to find same domain organization
Possible matches from various domain organizations
Limited diversity
Non-sensitive search
Enriched diversity
Sensitive search

Figure 2-8 Sensitivity problem of querying entire length of multi-domain protein in HMM-HMM search.

Coiled coils are structural units that are formed by anti-parallel two or three alpha helices. The amino-acid composition of these units shows a predictable pattern which is a combination of hydrophobic and hydrophilic amino acids. From sequence information only, coiled-coils can be easily predicted. COILS is an algorithm developed to predict these structural folds (Lupas et al. 1991).

Proteins can be cytoplasmic or membrane associated. Membrane-bound proteins have at least one region anchored within membrane. Transmembrane regions are composed of mostly hydrophobic residues. Though secondary structure of these regions are usually alpha helices, there are membrane proteins with beta barrel structures. Membrane-bound alpha helices can be computationally predicted with HMM approach. TMHMM is a popular tool to predict transmembrane alpha helices (Sonnhammer et al. 1998).

Signal peptides are short N-terminus regions of some proteins which are recognized by transporter proteins to transfer proteins to the subcellular location where they function. After translocation, signal peptidase cleaves this sequence. These regions are composed of hydrophobic residues, which in turn form an alpha-helix type structure. For this reason, signal peptides can be wrongly predicted as transmembrane helices. PHOBIUS is an algorithm to distinguish between signal peptides and transmembrane helices (Kall et al. 2007).

2.4.5 Domain coverage

Proteins can be defined and annotated based on the domain content. However, there are proteins that have not been found to be composed of any domain. Moreover, in multi-domain proteins, identifying only a single domain will not solve the problem of understanding the function of the entire protein. The orphan sequences have been referred as "dark matter" of protein universe (Rekapalli et al. 2012). Four potential reasons causing this dark matter of the proteins have been established: (i) errors in sequencing DNA, which would cause meaningless protein sequences and no similar sequences can be found; (ii) non-globular structures that are basically non-conserved disordered parts of proteins; (iii) inability to identify domains due to too diverged sequences by the course of evolution, which is a computational limitation; (iv) an encountered novel domain, which cannot be identified because of no similarity with the existing sequences (Levitt 2009). When inter-domain regions are subtracted orphan sequence ratio in protein universe was found to be ~40% in 2012. The relative size of uncovered amino acid sequences is shrinking every year. From April 2009 to December 2011, the Pfam coverage increased 3.4%. However, if the trend of domain coverage remains same, it will take more than 20 years to cover proteins in terms of domains. This fact shows a necessity of applying more sensitive domain search algorithms.

2.5 Human Genome and Genetic Diseases

2.5.1 Human Genome

Twenty years ago, an accurate estimate on the number of protein-coding genes in human genome couldn't be made. Though in 1960s 2,000,000 genes were estimated, in 1970s the upper limit for this count was determined as 40,000 (Cristianini and Hahn 2006). Recent work, although not precisely, shows that the human genome contains ~19,000 protein-coding genes which comprise only 1% of the human genome (Ezkurdia et al.

2014; Flicek et al. 2014). Overestimates on gene count were likely to be caused by the observation of the unique protein counts prior to the discovery of alternative splicing.

Humans and mice share majority of their genes with 85% average sequence identity in coding regions and their genomes are comparably similar in terms of content and sequence. Most of mice genes show noticeable phenotypes when knocked out and ~30% of them cause prenatal fatality. Therefore, each human gene is also suggested to be critical in survival. However, impacts of 52% of human genes have not been determined yet (Chong et al. 2015). Clearly, there is still a lot to be discovered on human genes, and thus health.

2.5.2 Variants

DNA polymorphism is a variation in nucleotide sequence that is common in population. An arbitrary cutoff percentage to categorize a DNA variation as polymorphism is 1%. It means that when a variation is observed in more than 1% of the population it is called polymorphism. DNA polymorphism and mutation are often misused interchangeably. Although every polymorphism can be defined as a mutation, not every mutation is polymorphism. Novel mutations that are population independent should not be called polymorphism. Most polymorphisms are benign. If they were damaging then they would have a fitness cost, which would have resulted in a reduced allele frequency. However, in biology there are many examples of DNA polymorphisms that are damaging at certain conditions. Also, combination of polymorphisms can determine the phenotype in health and disease. The most common type of polymorphism is a Single Nucleotide Polymorphism (SNP). SNPs can be within non-coding as well as coding DNA regions. Coding SNP types are basically categorized as: (i) Non-synonymous (ii) Synonymous and

(iii) Nonsense. Non-coding SNPs can affect transcription, especially if they are in functionally important regions such as the promoter, enhancer and silencer.

2.5.3 Sequencing Human Genes

Current sequencing technologies aim to be cheap, fast and high throughput. Since 2005, high throughput parallel sequencing approach is called Next Generation Sequencing (NGS). Today, NGS platforms can sequence a human genome within a day with a cost of a few thousand dollars.

2.5.3.1 Whole Genome Sequencing (WGS)

The process of deciphering the entire DNA code of an organism including mitochondria (or chloroplast in plants) is called whole genome sequencing (WGS). WGS has an important advantage as it provides information about not only genes but also any other DNA region. Thus, for complicated diseases caused by multiple coding and non-coding regions, WGS is the most appropriate choice due its comprehensive scan. However, it is expensive compared to the other specialized sequencing approaches. For testing specifically certain genotypes, WGS would be a waste of money and time. High coverage in sequencing is obtained with an extra cost. Because WGS is already expensive, coverage levels are kept at minimum levels in order not to increase the cost. The low coverage results in low confidence.

Human WGS results contain 6 billion (2 sets of chromosomes X 3 billion bases) nucleotides of information, in principle. Today, the analysis of the generated data is more challenging than sequencing. Although there is a number of robust algorithms available for processing the raw data, manual intervention is considered as a must to deduce biological meanings.

2.5.3.2 Whole Exome Sequencing (WES)

The entire set of exons in the genome is called exome. Whole exome sequencing focuses on the exome and ignores other DNA regions. Because many known genetic diseases are associated with coding regions, most clinicians are interested only in exons. In addition, because only ~1% of the human genome is composed of exons, WES is 100 times cheaper than WGS. It means that, 100 times more samples or coverage can be sequenced or achieved with the same cost. Like replication that DNA polymerases perform in cells, sequencing in laboratories is erroneous. The low cost of WES enables high coverage in DNA sequencing, which provides confidence in sequence analysis.

2.5.3.3 Targeted Sequencing

WGS and WES are usually used to identify the comprehensive list of variants, when a clinician suspects about a genetic case which has not been associated with a gene yet. However, in most cases DNA tests are performed to check if a certain gene/protein contains a mutation. In these cases only that part of the genome is targeted and sequenced. This approach is highly practical and preferable in terms of money, time and confidence.

2.5.4 Mendelian Diseases

A better understanding of diseases was probably the most encouraging motivation of the human genome project. Before the project started, the community expected to cure many diseases by understanding the causes with deciphered DNA code. However, sequenced human genome came with many unknowns and as of today, the causes of many genetic disorders are still unidentified. There are two types of genetic disorders classified based on their mechanism of action: Mendelian and complex disorders. Mendelian traits are consequences of single gene variations which can cause disorders such as sickle-cell disease and cystic fibrosis. The inheritance pattern of Mendelian traits is simple and traceable. In complex disorders such as diabetes and cardiovascular diseases, there are more than one gene involved. For multi-genic disorders it is challenging to understand the contribution of each gene involved in the disease. The primary aim of many researchers was to understand the simpler of the two types, Mendelian disorders. Mendelian traits are caused by loss of function, activity change, mislocalization of proteins. About 8% of people are diagnosed with a genetic disorder every year (Baird et al. 1988; Chong et al. 2015). Mendelian birth defects are the primary cause for the death under the age of one (Chong et al. 2015).

In clinics only well-established disorders can be diagnosed, whereas others cannot. Clinical diagnosis rate of Mendelian phenotypes is 50% (Chong et al. 2015). In children, diagnosis rate is as low as 11%. Besides, diagnosis success is another problem. There is a number of patients who are incorrectly diagnosed especially for rare diseases. Also, diagnosis periods are sometimes too long, which reduces the quality of life and results even in more severe conditions.

With recent developments in WGS and WES, the average number of discovered genes associated to a monogenic Mendelian phenotype has been increasing. WES has been the most widely used in diagnostics success compared to karyotyping and genomic hybridization. The WES related diagnose success rates of 25% - 30% depends on the recent discoveries on the association between genes and disorders. Thus, the future of the remaining 70% - 75% percent undiagnosed genetic disorders will depend on the findings on the gene-phenotype associations. As of today, Online Mendelian Inheritance in Man (OMIM) (McKusick 2007) database contains 7998 phenotypes described to have Mendelian basis. For ~45% of the Mendelian phenotypes, the molecular basis and responsible genes are still unknown.

For autosomal dominant diseases, at least one parent should display the diseased phenotype. In autosomal recessive diseases, if the parents are healthy, the both must be carriers and there is 25% chance of inheriting both disease alleles to the child. Heterozygosity is inferred from the relative ratio of alleles in sequence reads. Normally 50% is expected for heterozygous conditions.

2.5.5 Orthologs and Paralogs in Disease

Most of Mendelian phenotypes are caused by mutations in coding regions as they would have direct effects on proteins which are the fundamental biological molecules in cellular processes (Cooper et al. 2010).

Evolution of genes mainly depends on two major mechanisms: duplication and loss. These two events work together to invent new genes. An existing gene is duplicated in cellular processes such as homologous recombination, retrotransposition, chromosomal and whole genome duplication, and replication slippage. Right after duplication, the new gene is likely to be identical to the original one in terms of sequence. However, since one of the two is not going to be critically necessary, one of them will likely accumulate mutations and “differentiate”. Before the newer duplicate diverges, it may function in the same way of the original one. In that short time frame, the original gene may accumulate some mutations which would be compensated by the newer duplicate. Therefore a gene duplication may cause the original gene to be slightly diverged from the optimum fitness. In evolution, the redundancy of genes is not tolerated over long periods of time and one copy (usually the newly generated) either gains a new function or gets lost. These two sequences are still called homologous as one is derived from another. However, they are not orthologous as they are not resultant of a speciation event and they often have different biological functions. These genes are called paralogs and they are frequently associated with different phenotypes. For Mendelian diseases, it has been found that only one of the paralogous genes is associated with a disease **(Figure 2-9)** (Dickerson and Robertson 2012). Moreover, most of the Mendelian diseases are associated with the genes that have duplication history, which is likely to be caused by the slight divergence from the optimum fitness of the original gene upon duplication.

2.5.6 *In silico* Variant Assessment

Each human genome is estimated to contain 24,000-26,000 coding SNPs. Though substantial portion of the SNPs is synonymous, the non-synonymous part has a potential of changing the function of the associated protein. Each human genome contains 250-300 loss-of-function variations affecting the function of the protein which in turn may cause a disorder mostly at homozygotic or compound-heterozygotic state (1000-Genomes-Project-Consortium 2010). This is how rare genetic disorders arise. Although there are biochemical assays for well-known genetic disorders, not all of them are practical to be applied. Moreover, for cases where there are multiple suspects, testing each case is time consuming, financially unfavorable and labor intensive. Additionally,

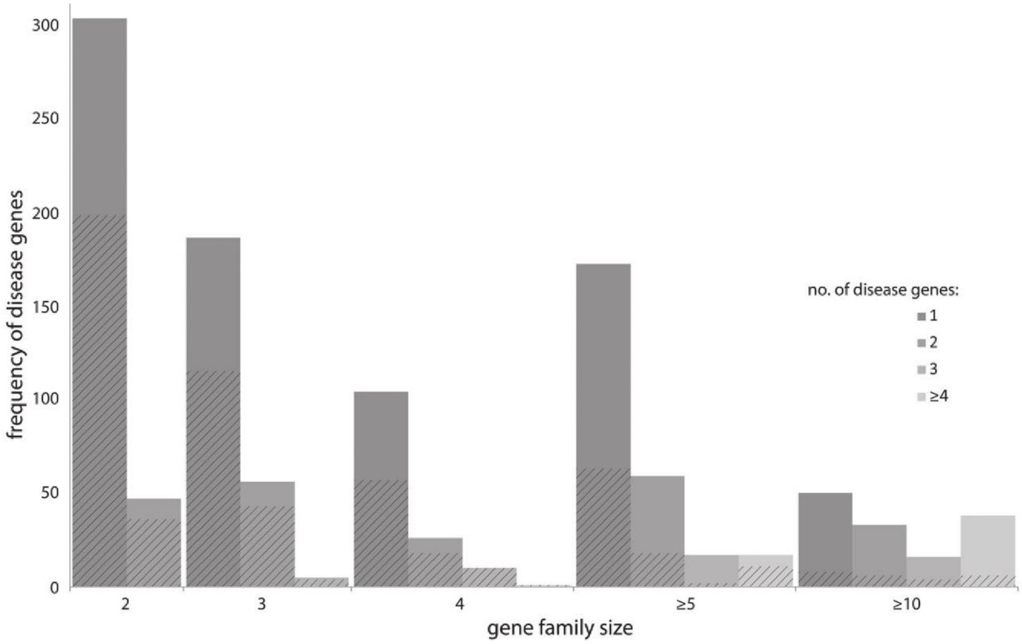


Figure 2-9 Gene family members are not involved in diseases equivalently.

Frequencies of disease genes from different sizes of gene families. Copied from (Dickerson and Robertson 2012)

biochemical test-based diagnostic decisions are made only for well-established disorders. For the remaining genetic disorders, a high throughput screening is needed. For these reasons, molecular testing has started to be widely applied in clinics. DNA sequencing is a cheap and standard method, and can play a vital role in diagnoses of some cases. However, the analysis part can be challenging especially if the variant of interest is unusual. When encountering a novel non-synonymous mutation, the question that clinicians asks is: “Does the mutation affect the protein function?” To answer this question, the only easy way is to computationally predict the potential effect of the variant on the function.

There is a number of variant assessment tools **(Table 2-2)**. Each of them has its own approach to evaluate the variants based on several features such as evolutionary history, structural and physiochemical importance of the substitutions. Physiochemical property change in a substitution can only be a contributing parameter, as it is not accepted to be standard in every case of molecular interaction between amino acids. Structures on the other hand can be useful, because biological function of proteins depends on their structural stability. However, structures of only 36% of human proteins have been solved so far. For this reason, evolutionary information set the basics of the algorithms assessing non-synonymous variants. Almost all of the protein-coding genes in human are conserved in vertebrate lineage (Aparicio et al. 2002; Mouse Genome Sequencing et al. 2002). The evolutionary depth for human genes should allow for the observation of benign mutations to be present in wild type genomes of other organisms, which is then translated to human variant interpretation. Therefore, whether a site is conserved or not has a substantial contribution in decision making about the risk of a variant. The current tools select a subset of related sequences by eliminating too close and too distant sequences and they do not consider the phylogenetic relationship between them.

Each of these automated tools uses their own approaches and as a result they lead to different predictions. The ratio of overlapping predictions for rare and novel variants between tools is fairly small (Chun and Fay 2009). So, a single tool is insufficient to confidently categorize variants. If a stringent categorization is desired, agreement between tools can be used as proposed in Chun et al. However, agreed results from multiple tools clearly reduces the specificity. For instance, in a research performed by

Table 2-2 Tools predicting effects of missense mutations on proteins. Modified from Richards et al.

Name	Reference	Basis
ConSurf	(Ashkenazy et al. 2010)	Evolutionary conservation
FATHMM	(Shihab et al. 2013)	Evolutionary conservation
MutationAssessor	(Reva et al. 2011)	Evolutionary conservation
PANTHER	(Thomas and Kejariwal 2004)	Evolutionary conservation
PhD-SNP	(Capriotti et al. 2006)	Evolutionary conservation
SIFT	(Ng and Henikoff 2003)	Evolutionary conservation
SNAP2	(Hecht et al. 2015)	Evolutionary conservation and predicted structural information
SNPs&GO	(Calabrese et al. 2009)	Protein structure/function
Align GVGD	(Mathe et al. 2006)	Protein structure/function and evolutionary conservation
MAPP	(Stone and Sidow 2005)	Protein structure/function and evolutionary conservation
MutationTaster	(Schwarz et al. 2010)	Protein structure/function and evolutionary conservation
MutPred	(Li et al. 2009)	Protein structure/function and evolutionary conservation
PolyPhen-2	(Adzhubei et al. 2013)	Protein structure/function and evolutionary conservation
PROVEAN	(Choi and Chan 2015)	Alignment and measurement of similarity between variant sequence
nsSNPAnalyzer	(Bao et al. 2005)	Multiple sequence alignment and protein structure analysis
Condel	(Gonzalez-Perez and Lopez-Bigas 2011)	Combines SIFT, PolyPhen-2, and MutationAssessor
CADD	(Kircher et al. 2014)	Contrasts annotations of fixed/nearly fixed derived alleles in humans with simulated variants

Chun et al., SIFT predicts 53% of the studied variants as damaging **(Figure 2-10)**. However, three tools agreed on only 7%. If the agreement of at least two tools criterion is used, 28% of variants are predicted as damaging. For this case, sensitivity reduces almost by half by using consensus of two tools. To sum up, neither a single tool nor a combination of tools is accurate enough to be confident and sensitive in predicting the damage of protein variants. This is one of the problems that clinicians are looking forward to having it solved.

To conclude, with the easiness of sequencing, prognosis and diagnosis of human genetic diseases should be made by testing DNA. However, a new challenge in molecular biology is now to analyze the sequence data rather than retrieving them. Because evolution primarily takes effect at the protein level, protein based predictions are preferred by the scientific community. Around 19,000 protein coding human genes can only be targeted by a robust approach. Automated tools, however, lack accuracy. Though manual analysis on proteins cannot be applied on large-scale, case by case, it has a potential to result in more accurate predictions in experts' hands, which in turn would be beneficial for the human health.

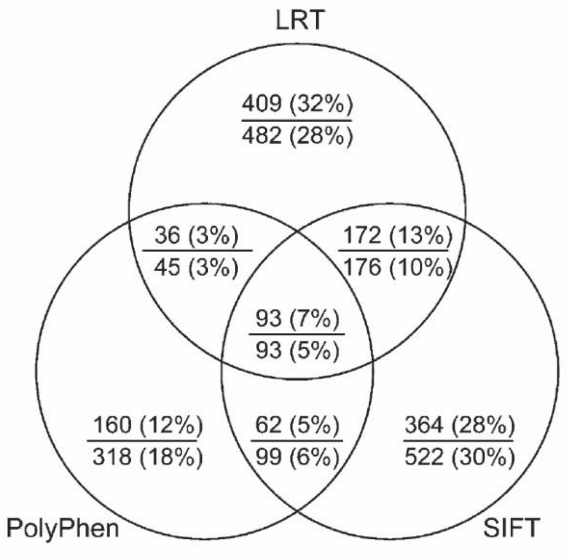


Figure 2-10 Inconsistent results produced by three different methods. Predictions made by three methods. Numbers below and above are variants in the Venter genome for the complete set and subset respectively.

Copied from Chun and Fay, 2009.

2.6 References

1000-Genomes-Project-Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061-1073.

Adzhubei I, Jordan DM, Sunyaev SR. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**: Unit7 20.

Altschul SF. 1991. Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol* **219**: 555-565.

Altschul SF. 1993. A protein alignment scoring system sensitive at all evolutionary distances. *J Mol Evol* **36**: 290-300.

Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG. 2015. Investigating Protein Structure and Evolution with SCOP2. *Curr Protoc Bioinformatics* **49**: 1 26 21-21 26 21.

Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A et al. 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* **297**: 1301-1310.

Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. 2010. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res* **38**: W529-533.

Baird PA, Anderson TW, Newcombe HB, Lowry RB. 1988. Genetic disorders in children and young adults: a population study. *Am J Hum Genet* **42**: 677-693.

Bao L, Zhou M, Cui Y. 2005. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Res* **33**: W480-482.

Bateman A, Finn RD. 2007. SCOOP: a simple method for identification of novel protein superfamily relationships. *Bioinformatics* **23**: 809-814.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res* **28**: 235-242.

Brunak S, Danchin A, Hattori M, Nakamura H, Shinozaki K, Matise T, Preuss D. 2002. Nucleotide sequence database policies. *Science* **298**: 1333.

Burki F, Shalchian-Tabrizi K, Minge M, Skjaeveland A, Nikolaev SI, Jakobsen KS, Pawlowski J. 2007. Phylogenomics reshuffles the eukaryotic supergroups. *PloS one* **2**: e790-e790.

Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. 2009. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum Mutat* **30**: 1237-1244.

Capriotti E, Calabrese R, Casadio R. 2006. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* **22**: 2729-2734.

Choi Y, Chan AP. 2015. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31**: 2745-2747.

Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, Harrell TM, McMillin MJ, Wiszniewski W, Gambin T et al. 2015. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am J Hum Genet* **97**: 199-215.

Chun S, Fay JC. 2009. Identification of deleterious mutations within three human genomes. *Genome Res* **19**: 1553-1561.

Coletta A, Pinney JW, Solís DY, Marsh J, Pettifer SR, Attwood TK. 2010. Low-complexity regions within protein sequences have position-dependent roles. *BMC systems biology* **4**: 43.

Cooper DN, Chen JM, Ball EV, Howells K, Mort M, Phillips AD, Chuzhanova N, Krawczak M, Kehrer-Sawatzki H, Stenson PD. 2010. Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics. *Hum Mutat* **31**: 631-655.

Cristianini N, Hahn MW. 2006. *Introduction to computational genomics: a case studies approach*. Cambridge University Press.

DePristo MA, Zilversmit MM, Hartl DL. 2006. On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins. *Gene* **378**: 19-30.

Di Roberto RB, Peisajovich SG. 2014. The role of domain shuffling in the evolution of signaling networks. *J Exp Zool B Mol Dev Evol* **322**: 65-72.

Dickerson JE, Robertson DL. 2012. On the origins of Mendelian disease genes in man: the impact of gene duplication. *Mol Biol Evol* **29**: 61-69.

Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, Vazquez J, Valencia A, Tress ML. 2014. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum Mol Genet* **23**: 5866-5878.

Felsenstein J. 1988. Phylogenies from molecular sequences: inference and reliability. *Annu Rev Genet* **22**: 521-565.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J et al. 2014. Pfam: the protein families database. *Nucleic Acids Res* **42**: D222-230.

Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, Bateman A, Eddy SR. 2015. HMMER web server: 2015 update. *Nucleic Acids Res* **43**: W30-38.

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S et al. 2014. Ensembl 2014. *Nucleic Acids Res* **42**: D749-755.

Galperin MY, Makarova KS, Wolf YI, Koonin EV. 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res* **43**: D261-269.

Geer LY, Domrachev M, Lipman DJ, Bryant SH. 2002. CDART: protein homology by domain architecture. *Genome Res* **12**: 1619-1623.

Gonzalez-Perez A, Lopez-Bigas N. 2011. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet* **88**: 440-449.

Gough J, Chothia C. 2002. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res* **30**: 268-272.

Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, Beck E. 2013. TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res* **41**: D387-395.

Hampl V, Hug L, Leigh JW, Dacks JB, Lang BF, Simpson AG, Roger AJ. 2009. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic “supergroups”. *Proceedings of the National Academy of Sciences* **106**: 3859-3864.

Hecht M, Bromberg Y, Rost B. 2015. Better prediction of functional effects for sequence variants. *BMC Genomics* **16 Suppl 8**: S1.

Heger A, Wilton CA, Sivakumar A, Holm L. 2005. ADDA: a domain database with global coverage of the protein universe. *Nucleic Acids Res* **33**: D188-191.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89**: 10915-10919.

Hildebrand A, Remmert M, Biegert A, Soding J. 2009. Fast and accurate automatic structure prediction with HHpred. *Proteins* **77 Suppl 9**: 128-132.

Horowitz S, Gorovsky MA. 1985. An unusual genetic code in nuclear genes of Tetrahymena. *Proc Natl Acad Sci U S A* **82**: 2452-2455.

Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. *Proceedings of the Royal Society of London B: Biological Sciences* **256**: 119-124.

Islam SA, Luo J, Sternberg MJ. 1995. Identification and analysis of domains in proteins. *Protein Engineering* **8**: 513-526.

Jones S, Stewart M, Michie A, Swindells MB, Orengo C, Thornton JM. 1998. Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Science* **7**: 233-242.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. *Mammalian protein metabolism* **3**: 21-132.

Kall L, Krogh A, Sonnhammer EL. 2007. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res* **35**: W429-432.

Kerfeld CA, Scott KM. 2011. Using BLAST to teach "E-value-tionary" concepts. *PLoS Biol* **9**: e1001014.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**: 310-315.

Koonin EV. 2010. The incredible expanding ancestor of eukaryotes. *Cell* **140**: 606-608.

Kummerfeld SK, Teichmann SA. 2005. Relative rates of gene fusion and fission in multi-domain proteins. *Trends Genet* **21**: 25-30.

Lees JG, Lee D, Studer RA, Dawson NL, Sillitoe I, Das S, Yeats C, Dessailly BH, Rentzsch R, Orengo CA. 2014. Gene3D: Multi-domain annotations for protein sequence and comparative genome analysis. *Nucleic Acids Res* **42**: D240-245.

Letunic I, Doerks T, Bork P. 2015. SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res* **43**: D257-260.

Levitt M. 2009. Nature of the protein universe. *Proc Natl Acad Sci U S A* **106**: 11079-11084.

Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. 2009. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* **25**: 2744-2750.

Lupas A, Van Dyke M, Stock J. 1991. Predicting coiled coils from protein sequences. *Science* **252**: 1162-1164.

Marchler-Bauer A, Bryant SH. 2004. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res* **32**: W327-331.

Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI et al. 2015. CDD: NCBI's conserved domain database. *Nucleic Acids Res* **43**: D222-226.

Mathe E, Olivier M, Kato S, Ishioka C, Hainaut P, Tavtigian SV. 2006. Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. *Nucleic Acids Res* **34**: 1317-1325.

McKusick VA. 2007. Mendelian Inheritance in Man and its online version, OMIM. *Am J Hum Genet* **80**: 588-604.

Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res* **41**: e121.

Moore AD, Bornberg-Bauer E. 2012. The dynamics and evolutionary potential of domain loss and emergence. *Mol Biol Evol* **29**: 787-796.

Mouse Genome Sequencing C Waterston RH Lindblad-Toh K Birney E Rogers J Abril JF Agarwal P Agarwala R Ainscough R Alexandersson M et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520-562.

Ng PC, Henikoff S. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**: 3812-3814.

Nikolskaya AN, Arighi CN, Huang H, Barker WC, Wu CH. 2006. PIRSF family classification system for protein functional and evolutionary analysis. *Evol Bioinform Online* **2**: 197-209.

Pruitt KD, Tatusova T, Brown GR, Maglott DR. 2012. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* **40**: D130-135.

Rekapalli B, Wuichet K, Peterson GD, Zhulin IB. 2012. Dynamics of domain coverage of the protein sequence universe. *BMC Genomics* **13**: 634.

Rekepalli BP. 2007. Automated Genome-Wide Protein Domain Exploration.

Remmert M, Biegert A, Hauser A, Soding J. 2012. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* **9**: 173-175.

Reva B, Antipin Y, Sander C. 2011. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* **39**: e118.

Salemi M, Vandamme A-M. 2003. *The phylogenetic handbook: a practical approach to DNA and protein phylogeny*. Cambridge University Press.

Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. 2010. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* **7**: 575-576.

Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, Day IN, Gaunt TR. 2013. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat* **34**: 57-65.

Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, Dawson NL, Furnham N, Laskowski RA, Lee D, Lees JG et al. 2015. CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res* **43**: D376-381.

Soding J. 2005. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**: 951-960.

Sonnhammer EL, von Heijne G, Krogh A. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* **6**: 175-182.

Stone EA, Sidow A. 2005. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res* **15**: 978-986.

Thomas PD, Kejariwal A. 2004. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. *Proc Natl Acad Sci U S A* **101**: 15398-15403.

Ulrich LE, Zhulin IB. 2014. SeqDepot: streamlined database of biological sequences and precomputed features. *Bioinformatics* **30**: 295-297.

UniProt C. 2008. The universal protein resource (UniProt). *Nucleic Acids Res* **36**: D190-195.

UniProt C. 2015. UniProt: a hub for protein information. *Nucleic Acids Res* **43**: D204-212.

Wheelan SJ, Marchler-Bauer A, Bryant SH. 2000. Domain size distributions can predict domain boundaries. *Bioinformatics* **16**: 613-618.

Wootton JC. 1994. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem* **18**: 269-285.

CHAPTER 3. CDvist: A Webserver for Identification and Visualization of Conserved Domains in Protein Sequences

This chapter was originally published by Adebali et al. in Bioinformatics.

Ogun Adebali, Davi R. Ortega, and Igor B. Zhulin. "**CDvist: a webserver for identification and visualization of conserved domains in protein sequences.**" *Bioinformatics* 31.9 (2015): 1475-1477.

Author's contribution: O.A. and I.B.Z. designed and led the project; O.A. developed the algorithm and built the web server; O.A., D.R.O. and I.B.Z. wrote the manuscript.

3.1 Abstract

Summary

Identification of domains in protein sequences allows their assigning to biological functions. Several webservers exist for identification of protein domains using similarity searches against various databases of protein domain models. However, none of them provides comprehensive domain coverage while allowing bulk querying and their visualization schemes can be improved. To address these issues we developed CDvist (a comprehensive domain visualization tool), which combines the best available search algorithms and databases into a user friendly framework. First, a given protein sequence is matched to domain models using high specificity tools and only then unmatched segments are subjected to more sensitive algorithms resulting in a best possible comprehensive coverage. Bulk querying and rich visualization and download options provide improved functionality to domain architecture analysis.

Availability

Freely available on the web at http://cdvist.utk.edu

Contact

oadebali@vols.utk.edu or ijouline@utk.edu

3.2 Introduction

The identification of protein domains is a key feature of protein sequence analysis. Several databases, notably Pfam (Punta et al. 2012), SMART (Letunic et al. 2009), COG (Tatusov et al. 2003), CDD (Marchler-Bauer et al. 2013) and others, develop and maintain domain models. Searching tools such as RPS-BLAST (Marchler-Bauer et al. 2013), HMMER3 (Eddy 2011) and HHpred/HHsearch (Soding 2005; Hildebrand et al. 2009) are used to match sequences to domain models present in a given database. The size of the protein sequence database grows dramatically, whereas its coverage by precomputed domain models increases very slowly (Rekapalli et al. 2012). Consequently, sensitive domain searches of sequences in bulk are necessary to improve computational coverage of the current and future protein sequence space. Despite the overwhelming success of the current state-of-the-art domain searching resources, three areas require further improvements: i) combining tools with high specificity and tools with high sensitivity in a single framework, ii) multiple query searches using highly sensitive (e.g. profile-to-profile) methods, iii) visualization of most relevant information in a responsive and interactive way.

To address these issues, we have developed the Comprehensive Domain Visualization Tool (CDvist), a domain searching webserver specialized in maximizing domain coverage of multi-domain protein sequences with emphasis on visualization.

3.3 Implementation and Features

Users submit protein sequences in FASTA format and each sequence is processed independently of each other on individual linux cluster nodes. Up to 500 queries per request are supported. The following domain search methods are implemented in CDvist: HMMER3 (Eddy 2011), RPS-BLAST (Schaffer et al. 1999; Marchler-Bauer et al. 2013), HHSEARCH (Soding 2005), and HHBLITS-HHSEARCH (Remmert et al. 2012). Transmembrane regions are predicted by either TMHMM (Sonnhammer et al. 1998) or Phobius (Kall et al. 2007). Low complexity and coiled coil regions are predicted by SEG (Wootton 1994) and Coils (Lupas 1996) respectively. To improve domain coverage, rather than using the entire sequence, CDvist iteratively identifies regions without significant domain match (orphan segments) and submits each one of them to similarity search against a user-determined sequence of databases until the entire protein sequence is covered or all databases have been searched **(Figure 3-1)**. The key principle of this process is that tools that have high specificity – HMMER against Pfam and RPS-BLAST against CDD – are used first. Only then, the sequence segments that were not confidently matched to any model are used to build profiles and subjected to more sensitive profile-profile searches by HHsearch. Each algorithm can be turned on/off and the order of databases, and their significance thresholds, can be altered. This flexibility enables users to tailor the overall process for their specific purposes. Optional ‘domain split’ function splits the matched domain model if there is a considerable unaligned query region (5% by default) in the query-model alignment. This unaligned region is considered as an orphan segment and is used in the next run to search for potential domains.

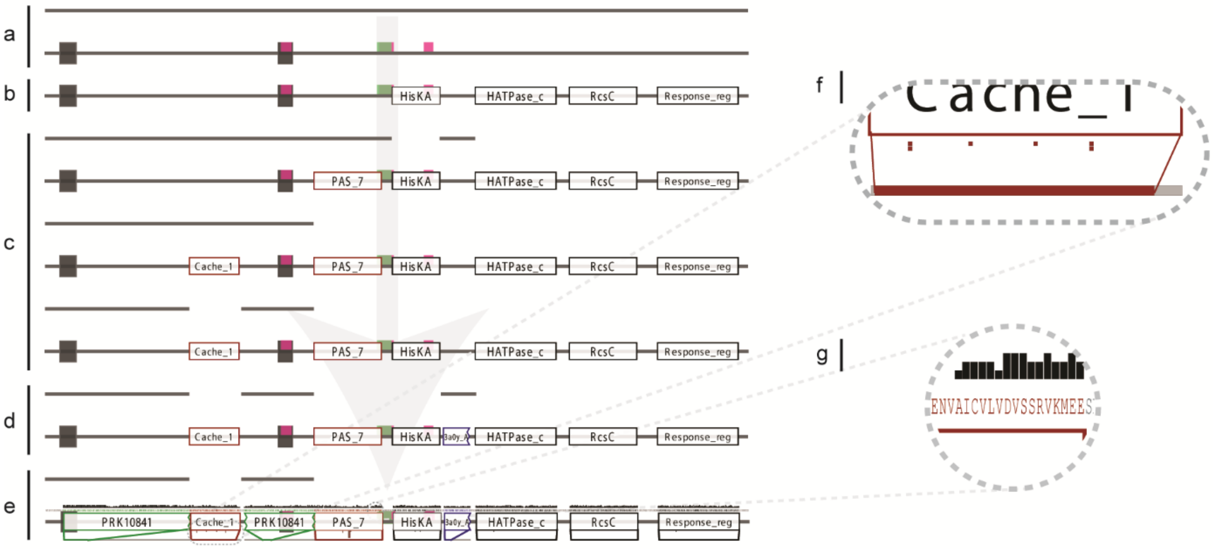


Figure 3-1 Workflow and visualization example.

a) Primary sequence is used as input and transmembrane (gray), low complexity (magenta) and coiled-coil (green) regions are predicted. **b)** HMMER3 scan against Pfam database is executed and first domain architecture is built. **c-e**) HHblits followed by HHsearch is executed against **c)** Pfam, **d)** PDB and **e)** CDD databases. **f)** Domain coverage option: gray background represents the whole length of model whereas red bar displays the portion of the model that aligns with the query. Square points represent the domain positions that do not align with the query. **g)** Alignment option. Sequence is displayed to scale, and each bar stands for alignment quality at that position. The absence of the bar at a given position indicates gap in the alignment on query side.

A custom built JavaScript module powers the visualization on the client side with images in vector format (SVG) that are practical to edit, export as PDF and produce figures of publication quality. Results are displayed asynchronously for each query sequence submitted, which also allows the user to interact with the data before the completion of the entire request. Domain coverage bar provides information on what portion of the matched domain model is represented on the query sequence **(Figure 3-1f)**. Alignment quality is represented as vertical bar for each position of the alignment. Gaps in the alignment indicate that the corresponding part of the query is not aligned with the model **(Figure 3-1g)**. Scaled sequence information is mapped on the domain architecture, which is easily retrievable by zooming in on the browser. Drag feature allows user to align desired parts of batch data for further analysis. All this information is hosted in our webserver for over a week with a unique URL. Alternatively, the user can retrieve the HTML file to control the interactive feature visualizations locally on a web browser. JSON formatted files containing the information used to draw the graphics in the website are available for not only for each individual sequences but also for the entire input set as a single file. Finally, the log files for each run are available, which display the raw output of the whole process. Logs provide extra information on less significant hits which are not displayed visually. The databases are updated immediately upon their release.

3.4 Discussion

CDvist is designed to provide maximum domain coverage in protein sequences by bundling the best current domain search tools into a pipeline that exhaustively searches through a series of domain databases in an iterative fashion. This methodology yields the most comprehensive domain architecture for a given protein sequence. Rich visualization, download options and linear speed-up for bulk queries should be appealing to both biologists and bioinformaticians. This webserver would be especially useful for multi-domain proteins with rare or unique domain architectures and those prone to domain swap, where whole sequence similarity searches often yield uninformative and misleading results (Iyer et al. 2001).

3.5 Acknowledgments

Funding

This work was supported by NIH grant GM072285 (to I.B.Z.). O.A. received support from the UT-ORNL Graduate Program of Genome Science and Technology.

3.6 References

Eddy SR. 2011. Accelerated Profile HMM Searches. *PLoS computational biology* **7**: e1002195.

Hildebrand A, Remmert M, Biegert A, Soding J. 2009. Fast and accurate automatic structure prediction with HHpred. *Proteins* **77 Suppl 9**: 128-132.

Iyer LM, Aravind L, Bork P, Hofmann K, Mushegian AR, Zhulin IB, Koonin EV. 2001. Quod erat demonstrandum? The mystery of experimental validation of apparently erroneous computational analyses of protein sequences. *Genome biology* **2**: RESEARCH0051.

Kall L, Krogh A, Sonnhammer EL. 2007. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic acids research* **35**: W429-432.

Letunic I, Doerks T, Bork P. 2009. SMART 6: recent updates and new developments. *Nucleic acids research* **37**: D229-232.

Lupas A. 1996. Coiled coils: new structures and new functions. *Trends in biochemical sciences* **21**: 375-382.

Marchler-Bauer A, Zheng C, Chitsaz F, Derbyshire MK, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Lanczycki CJ et al. 2013. CDD: conserved domains and protein three-dimensional structure. *Nucleic acids research* **41**: D348-352.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J et al. 2012. The Pfam protein families database. *Nucleic acids research* **40**: D290-301.

Rekapalli B, Wuichet K, Peterson GD, Zhulin IB. 2012. Dynamics of domain coverage of the protein sequence universe. *BMC genomics* **13**: 634.

Remmert M, Biegert A, Hauser A, Soding J. 2012. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods* **9**: 173-175.

Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF. 1999. IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics* **15**: 1000-1011.

Soding J. 2005. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**: 951-960.

Sonnhammer EL, von Heijne G, Krogh A. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology* **6**: 175-182.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN et al. 2003. The COG database: an updated version includes eukaryotes. *BMC bioinformatics* **4**: 41.

Wootton JC. 1994. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Computers & chemistry* **18**: 269-285.

CHAPTER 4. Aquerium: A Web-based Platform for Phylogenetic Profiling of Protein Domain Architectures

This chapter was taken from a manuscript in preparation.

Ogun Adebali and Igor B. Zhulin, **Aquerium: a web-based platform for phylogenetic profiling of protein domain architectures.** Manuscript in preparation.

Author's contribution: O.A. designed the project, developed the tool and built the web server; O.A. and I.B.Z. discussed the results and wrote the manuscript.

4.1 Abstract

Gene duplication and loss are major driving forces in evolution. While many important genomic resources provide information on gene presence, there are no tools for retrieving information on gene absence. Here, we present Aquerium, a platform for visualizing genomic presence and absence of biomolecules with a focus on protein domain architectures. The webserver offers advanced domain organization querying against the database of pre-computed domains for ~26000 organisms and it can be utilized for identification of evolutionary events, such as fusion, disassociation, duplication and shuffling of protein domains. The tool also provides alternative inputs of custom entries or BLAST results for visualization. Aquerium is available at http://aquerium.utk.edu.

4.2 Introduction

Phylogenetic profiling is a method to detect functionally or physically interacting proteins by inferring their co-presence/absence in hierarchically clustered species (Skunca and Dessimoz 2015). If genes are gained or lost together, it is likely that their products participate in the same biological pathway, meaning that they interact functionally. The method was first described by Pellegrini *et al.* who investigated the coevolution patterns of *Escherichia coli* genes (Pellegrini et al. 1999). They demonstrated that gene groups that have similar occurrence profiles tend to be involved in the same pathways. Consequently, in addition to discovering protein-protein interactions, phylogenetic profiling can be used for protein function prediction. There has been a number of successful applications of this context-based method complemented by homology-based and experimental approaches (Kensche et al. 2008).

Homology is inferred through sequence-based similarity searches. Domain organization comparisons can also be used to infer homology and to identify protein families. Domains, defined as minimal structural and functional building blocks of proteins, are capable of folding autonomously and evolving independently. Single domain proteins (SDPs) were likely dominant in the early stage of life, whereas multi-domain proteins (MDPs) are enriched with the complexity of organisms (Di Roberto and Peisajovich 2014). In SDPs the domain itself functions alone while in MDPs domains work in collaboration to perform the protein function.Domains can exist in various arrangements in a protein and this flexibility enriches the diversity of protein families. The complexity of MDPs can be attributed to the evolutionary dynamic nature of domains. The evolutionary events, such as domain innovation, loss, duplication, fusion, disassociation and shuffling enable proteins and eventually organisms to adapt to their environment (Kummerfeld and Teichmann 2005). Particularly, domain shuffling, rather than *de novo* inventions from disordered sequences is the major evolutionary event to generate novel proteins (Di Roberto and Peisajovich 2014). It was suggested that the total number of unique protein domains decreased in the course of eukaryotic evolution. For instance, last eukaryotic common ancestor (LECA) had a larger unique domain pool than any of the current species (Zmasek and Godzik 2011). Also, in mammals, a drop in the domain pool has been observed compared to the ancestral repository (Zmasek and Godzik 2011). These observations suggest that reusing protein domains in various modifications and rearrangements drives protein evolution.

More complex organisms have relatively more complex MDPs (Das et al. 1997; Wolf et al. 1999). This correlation may explain why gene number does not increase with organismal complexity (Koonin et al. 2002). Therefore, in order to understand complex networks, it is important to investigate the function of domains and how they collectively work together. Inferring the evolutionary relationships between domains is critically important in order to identify their functions and interactions.

Domains in protein sequences can be identified computationally, e.g. using. HMM (Hidden Markov Model) profiles. Pfam (Protein Families) database is a large collection of HMMs and underlying tools, which is one of the most popular resources for identifying protein domains (Finn et al. 2014). Pfam-A, a manually curated subset of the database, currently (version 28.0) contains 16230 domain models. Another HMM utilizing resource, TIGRFAM, contains models for many full-length proteins so that it provides an easy detection for protein families (Haft et al. 2013).

Biological networks diverge from their ancestor by protein or domain gaining/losing and domain shuffling. Such diversity patterns can be detected by comparative analysis of domain architectures. For this reason, retrieving the domain organization of interest and visualizing its taxonomic distribution are the crucial steps in understanding the functional relationships within networks. In addition to Pfam, several other tools, such as SMART (Letunic et al. 2015), CDART (Geer et al. 2002), DAhunter (Lee and Lee 2008) and

PfamAlyzer (Hollich and Sonnhammer 2007), provide domain architecture querying. These tools specialize in searching for protein homologies through similar domain architectures. However, none of the current resources allows advanced domain architecture querying while visualizing domain presence and absence in the phylogenomic context. Furthermore, none of the tools has an option to visualize genomic distributions of multiple queries at once.

To address these problems, we developed Aquerium (architecture querying podium), a tool enabling biologists and bioinformaticians to understand the domain-based evolutionary history of proteins.

4.3 Implementation

Genomes from the NCBI genomes database (as of 12th of December 2014) which also had assembly records in the NCBI assembly database were selected. GenBank records for each genome was retrieved from the NCBI Entrez Genome database (Gibney and Baxevanis 2011). We created a proteome collection for each genome. In order to manage isoforms in eukaryotes, each protein was categorized under the gene identifier that it is coded by. If a gene has at least one protein isoform matching the query, the tool returns true. If several isoforms match with a query, only one of them is taken into account in order to eliminate redundancy in count number. SeqDepot database (Ulrich and Zhulin 2014) was used locally to retrieve the pre-computed domain architectures from Pfam versions 27.0 and 28.0 and TIGRFAM versions 14 and 15. Local SeqDepot database was updated by running HMMER3 (Mistry et al. 2013) searches against domain databases for uncovered proteins.

The NCBI taxonomy database (Federhen 2012) was used to build the tree. In addition to eight major taxonomic ranks, we also included the five eukaryotic supergroups (Koonin 2010). Protein GI number to taxonomic id mapping was performed using the local database that was retrieved from NCBI taxonomy ftp source (Federhen 2012) and daily updated. The resulting taxonomic tree can be visualized by using two sets of genomes: species-representative and full sets. Species-representative set (4934 genomes) was built by selecting only one representative for strains determined by their species-level taxonomic ids. The genomes with the largest number of genes among strains were selected as representative. The full set was composed of 26618 organisms.

The data has been organized in a document based MongoDB database. Custom Python3 scripts were developed for searching the database. JavaScript was implemented in HTML5 to visualize the results. The final figure is drawn in Scalable Vector Graphics (SVG).

4.4 Features

Advanced domain architecture querying

MDPs have various domain arrangements. In some proteins, the domain order is conserved, whereas other proteins are subjected to domain shuffling, duplication and loss. Diverged domain architectures might be indicators of modified or adapted function. For these reasons, it is important to enable extensive architecture querying.

Aquerium allows users to select the domain of interest, called “key domain”, to initialize the search. This field is mandatory and the algorithm will retrieve proteins that have at least one key domain. In the query page, a condition (“if” statement in Python syntax) can be specified to customize the query in terms of domain content and organization. This condition is used for enriched querying in which presence and absence of other domains can be examined. Moreover, the order of the domains, from N- to C-terminus can be specified and only proteins satisfying the given condition are retrieved. Specifying a condition is not necessary if the user is interested only in the presence and absence of the key domain. Domain search can be performed on species-representative and full sets and these sets can be filtered based on taxonomic units.

Visualization

Species are clustered based on their taxonomic ranks and represented as a sunburst tree on which each taxonomic class is drawn as an arc. The length of arcs scales to represent the number of species which are eventual descendants of the node. On the tree, there are nine taxonomic layers representing the major taxonomic ranks and supergroups for eukaryotes (Koonin 2010). After taxonomic ranks, each outer ring represents the requested query. If there is any match in the corresponding genome, there will be a colored flag aligning with the organism on an outer circle.

In the “zoom” mode each taxonomic node, represented by an arc, is zoomable on click. The sunburst is redrawn and shows only the selected node and its children in a circular layout. Extensive coloring options are offered on the fly allowing to produce publication-quality figures. The coloring of flags can also be performed as a heatmap depending on the quantity of each flag. Multiple layers can be visualized on the same tree. Users can visualize up to 10 outer layers on the same tree. In the “Arc” mode, clicking on a node will redirect the user to another webpage where they can visualize the associated organisms and the domain architectures on a collapsable tree layout.

Data export

The sunburst tree can be exported in scalable vector graphics (SVG). The compiled data can be downloaded in semicolumn separated file (CSV) format, which includes the taxonomic identifiers and the number of occurrences for each organism. JSON file containing taxonomically classified organism information is also available to retrieve. Moreover, protein sequence (in FASTA format) download option for desired taxonomic unit is available.

Custom input for visualization

In addition to protein domains, the sunburst tree can be produced with any other types of genomic data. Users can input a custom table containing NCBI taxonomic id followed by numeric or binary occurrence of profiles in CSV format and visualize the results. Up to 10 flag layers can be visualized in a single request.

Aquerium web server also offers visualizing blastp results on the tree. Users must download xml version of the BLASTp (Boratyn et al. 2013) results from NCBI and upload it to the relevant link on the Aquerium web page. Filtering the blast hits are possible by setting up thresholds for e-value, query and subject coverage. Additionally, protein GI number list can also be used as an input.

4.5 Illustrations

In order to exemplify Aquerium performance, we presented two independent test cases which show potential applications to similar problems.

Identification of a domain fusion event. Amino acid kinase (AA_kinase) and Aldehyde dehydrogenase (Aldedh) domain families are universal and seen in all domains of life with minor absences in few parasitic clades. These domains usually comprise a single domain protein, such as *E. coli* glutamate-5-kinase and γ-glutamyl phosphate reductase. Human δ-1-pyrroline-5-carboxylate synthetase has evolved as a fusion product of AA_kinase and Aldedh domains (Marcotte et al. 1999). Figure 4-1 shows the presences and absences of these domains. The outmost layer shows the occurrence of these two domains together in a single protein. In all supergroups of eukaryotes, these two domains are fused. The observed pattern of inheritance suggests that the fusion of these domains has occurred in the common ancestor of eukaryotes, and the common ancestor of fungi lost it.

Coexisting proteins and abundance correlation. The signaling complex in bacterial chemotaxis, which has been conserved since the common bacterial ancestor (Wuichet and Zhulin 2010), consists of MCPs (chemoreceptors), CheA (a kinase) and CheW (an adaptor). These three proteins are found together in 98% of genomes that encode chemotaxis genes (Wuichet and Zhulin 2010). Figure 4-2 shows the phlyetic distributions of these three proteins. Satisfactorily, in the vast majority of cases, all three proteins are either present or absent in genomes indicating the presence or absence of chemotaxis as a cellular function. This test case serves as a control for true negatives. Relative abundances of these proteins in genomes (some genomes have several different types of the signaling complex encoded by different sets of genes) also correlate. This is visualized using the heatmap option revealing the number of hits for each organism. Increased abundance is shown by a change of color intensity from light to dark.

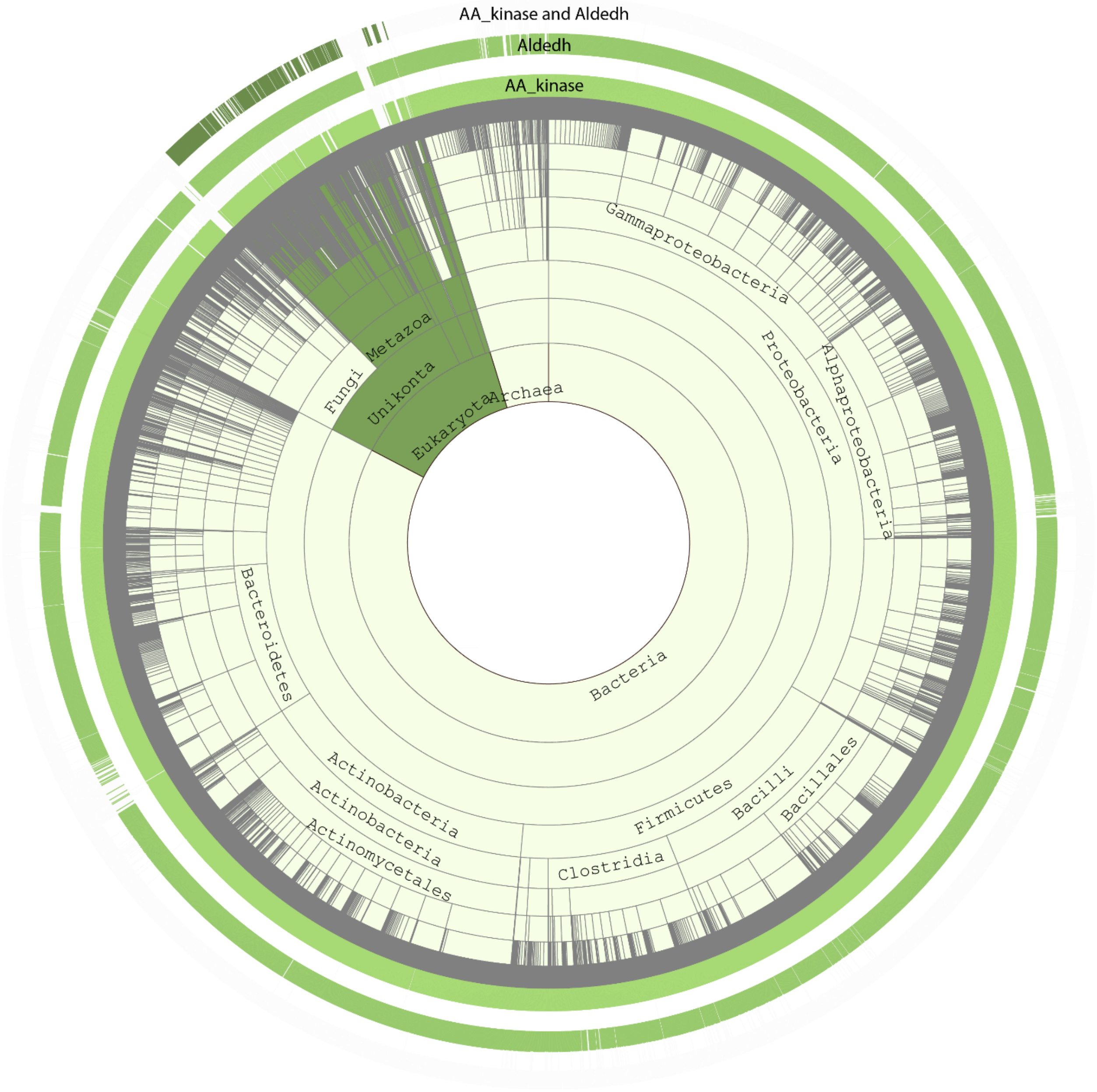


Figure 4-1 Illustration of a domain fusion event.

Fused proteins containing both Aldedh and AA_kinase domains are found in all represented eukaryotic supergroups, suggesting that the fusion occurred in the last eukaryotic common ancestor (LECA).

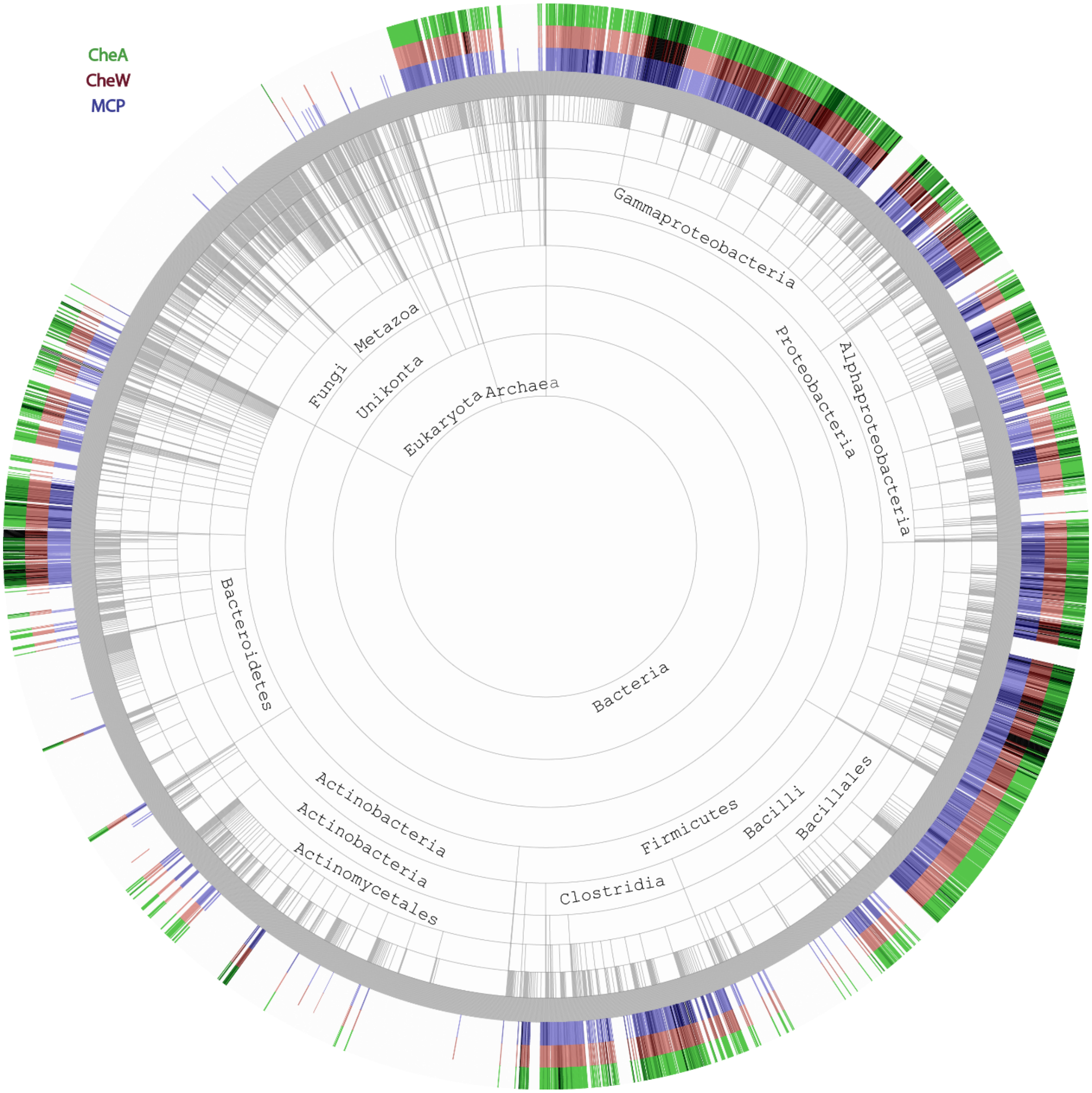


Figure 4-2 Interacting proteins coevolve.

Chemotaxis proteins MCP, CheA and CheW are known to be interacting with each other. They show similar patterns of not only occurrence but also relative abundance.

4.6 Discussion

The presence of genetic material in a genome is almost never questioned except for the possibility of contamination. On the other hand, the absence is always questioned and negative information should be treated cautiously. Being confident about the absence of particular genes/proteins/domains in genomes is challenging for two main reasons: (i) genomes may be incomplete, erronenus or contaminated and (ii) genes may not be identified due to computational limitations. However, the absence of two or more genes/proteins/domains that is consistently observed in independent samplings strongly suggests that the absence is true (Figure 2). Independent co-evolution can be identified by large-scale analyzes; as the number of samples increases, the likelihood of finding independent cases also increases.

Aquerium enables exploring a variety of phenomena in a genomic context, ranging from evolution of individual domains to inferring potential protein-protein interactions, by placing a nearly equal weight on the presence and the absence of genomic entities, such as genes, proteins and their domains.Thus, this tool is expected to be useful to many biologists working within the genomic landscape.

4.7 Acknowledgments

This work was supported in part by the National Institutes of Health grants GM072285 and DE024463.

4.8 References

Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, Madden TL, Matten WT, McGinnis SD, Merezhuk Y et al. 2013. BLAST: a more efficient report with usability improvements. *Nucleic acids research* **41**: W29-33.

Das S, Yu L, Gaitatzes C, Rogers R, Freeman J, Bienkowska J, Adams RM, Smith TF, Lindelien J. 1997. Biology's new Rosetta stone. *Nature* **385**: 29-30.

Di Roberto RB, Peisajovich SG. 2014. The role of domain shuffling in the evolution of signaling networks. *Journal of experimental zoology Part B, Molecular and developmental evolution* **322**: 65-72.

Federhen S. 2012. The NCBI Taxonomy database. *Nucleic acids research* **40**: D136-143.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J et al. 2014. Pfam: the protein families database. *Nucleic acids research* **42**: D222-230.

Geer LY, Domrachev M, Lipman DJ, Bryant SH. 2002. CDART: protein homology by domain architecture. *Genome research* **12**: 1619-1623.

Gibney G, Baxevanis AD. 2011. Searching NCBI Databases Using Entrez. *Curr Protoc Hum Genet* **Chapter 6**: Unit6 10.

Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, Beck E. 2013. TIGRFAMs and Genome Properties in 2013. *Nucleic acids research* **41**: D387-395.

Hollich V, Sonnhammer EL. 2007. PfamAlyzer: domain-centric homology search. *Bioinformatics* **23**: 3382-3383.

Kensche PR, van Noort V, Dutilh BE, Huynen MA. 2008. Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. *Journal of the Royal Society, Interface / the Royal Society* **5**: 151-170.

Koonin EV. 2010. Preview. The incredible expanding ancestor of eukaryotes. *Cell* **140**: 606-608.

Koonin EV, Wolf YI, Karev GP. 2002. The structure of the protein universe and genome evolution. *Nature* **420**: 218-223.

Kummerfeld SK, Teichmann SA. 2005. Relative rates of gene fusion and fission in multi-domain proteins. *Trends in genetics : TIG* **21**: 25-30.

Lee B, Lee D. 2008. DAhunter: a web-based server that identifies homologous proteins by comparing domain architecture. *Nucleic acids research* **36**: W60-64.

Letunic I, Doerks T, Bork P. 2015. SMART: recent updates, new developments and status in 2015. *Nucleic acids research* **43**: D257-260.

Marcotte EM, Pellegrini M, Ng H-L, Rice DW, Yeates TO, Eisenberg D. 1999. Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**: 751-753.

Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic acids research* **41**: e121.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America* **96**: 4285-4288.

Skunca N, Dessimoz C. 2015. Phylogenetic profiling: how much input data is enough? *PloS one* **10**: e0114701.

Ulrich LE, Zhulin IB. 2014. SeqDepot: streamlined database of biological sequences and precomputed features. *Bioinformatics* **30**: 295-297.

Wolf YI, Brenner SE, Bash PA, Koonin EV. 1999. Distribution of protein folds in the three superkingdoms of life. *Genome research* **9**: 17-26.

Wuichet K, Zhulin IB. 2010. Origins and diversification of a complex signal transduction system in prokaryotes. *Sci Signal* **3**: ra50.

Zmasek CM, Godzik A. 2011. Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. *Genome biology* **12**: R4.

CHAPTER 5. Establishing Precise Evolutionary History of a Gene Improves Predicting Disease Causing Missense Mutations

This chapter was taken from a manuscript submitted.

Ogun Adebali, Alexander O. Reznik, Daniel S. Ory and Igor B. Zhulin, **Establishing Precise Evolutionary History of a Gene Improves Predicting Disease Causing Missense Mutations.** Manuscript submitted.

Author's contribution: O.A. and I.B.Z designed the research; O.A., A.O.R. and I.B.Z. performed the research; O.A., A.O.R., D.S.O. and I.B.Z. analyzed the data; O.A., D.S.O. and I.B.Z. wrote the manuscript.

5.1 Abstract

Predicting the phenotypic effects of mutations has become an important application in population genetics studies and clinical genetic diagnostics. Computational tools, such as PolyPhen and SIFT, utilize comparative genomics to evaluate the behavior of the variant over evolutionary time and assume that variants seen during the course of evolution are likely benign in humans. However, these tools do not take into account orthologous/paralogous relationships. Paralogs have dramatically different roles in Mendelian diseases. For example, while 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. Here we show that by removing the *NPC1* paralogs from the analysis we can improve the overall performance of categorizing damaging and benign single amino acid substitutions. We anticipate that this approach will improve the interpretation of variants in other genetic diseases as well.

5.2 Introduction

With the revolutionary developments in sequencing technologies (Katsanis and Katsanis 2013), molecular testing is now widely used to confirm or support clinical diagnosis. Being cheap, fast and accurate, DNA sequencing is a promising method for prognosis, diagnosis, personalized therapeutics and identifying unknown cause in genetic disorders (Ng et al. 2010; Chang and Li 2013; Katsanis and Katsanis 2013). There are several approaches to evaluate the effect of a variant: (i) evidence-based, (ii) frequency-based, (iii) functional (variants with obviously drastic consequences such as nonsense and frameshift mutations), and (iv) predictive (Oliver et al. 2015). Being knowledge-based, the first three approaches are often successful in determining the effects of variants. However, they are limited when it comes to the variants of unknown significance (Katsanis and Katsanis 2013). For novel variants, which comprise the vast majority of coding variation (Tennessen et al. 2012), *in silico* prediction is a quick way is to estimate potential consequences. There is a number of computational tools, such as PolyPhen (Adzhubei et al. 2010) and SIFT (Ng and Henikoff 2003), that are frequently used to evaluate genetic variations not only in research laboratories, but also in clinical practices. However, they are not yet at the level of desired performance in terms of sensitivity and specificity, even for well-studied monogenic Mendelian diseases (Jordan et al. 2010; Sunyaev 2012; Oliver et al. 2015). Therefore, there is still a need for improvement in computational prediction of variant effects (Oliver et al. 2015).

Current tools that are automated, fast and applicable to all human protein-coding genes consider the following key parameters: sequence conservation, structural constraints and physiochemical properties of amino acids. Risk estimation is largely dependent on the molecular conservation, which is inferred from comparative sequence analysis (Ng and Henikoff 2006). The motivation behind using molecular conservation as a key estimate is the fact that deleterious mutations cause a reduction of evolutionary fitness; therefore, they are not selected for and are not observed in homologs in other organisms (Jordan et al. 2010). In order to identify homologous sequences in other organisms, current tools employ automated sequence similarity searches followed by clustering. Consequently, sets of similar sequences that are used in the downstream analysis usually include both orthologs and paralogs (Adzhubei et al. 2010). This approach is based on the argument that disease-causing substitutions far more often affect protein structure than function (Wang and Moult 2001), and while paralogous proteins may have a slightly different function, their structure is fully conserved.

However, recent studies revealed that the roles of paralogous genes in disease and health are different. In most of the cases of Mendelian diseases, among gene family members, only one gene is associated with the disease, while others do not have any role in that particular disorder (Dickerson and Robertson 2012). In 87% percent of the gene pairs, only one pair is associated with disease, and this trend is observed in gene families with more than two members. Duplication of genetic material is the primary source of new protein-coding elements rather than *de novo* invention. Once a gene is duplicated, purifying selection pressure on one of the copies is relaxed and that gene becomes more prone to accumulating mutations. This divergence can lead to sub-functionalization, neo-functionalization or non-functionalization of the paralogous gene (Lynch and Conery 2000) often resulting in their different roles of paralogs in disease.

This precise pattern is observed in Niemann-Pick disease type C (NP-C), which is a neurovisceral lysosomal lipid storage disease with an incidence of 1:100,000 (Vanier 2010; Patterson et al. 2012; Jahnova et al. 2014). NP-C is inherited in autosomal recessive pattern and caused by mutations in either *NPC1* or *NPC2* genes (Vanier 2015). NPC1 and NPC2 proteins work in concert to transport cholesterol from the endosomal/lysosomal compartment (Sleat et al. 2004; Vanier 2010). Homozygous loss of function in either protein perturbs lipid homeostasis, specifically by causing sterol and sphingolipid accumulation in the late endosomal/lysosomal (LE/L) compartment of cells, which results in pathogenicity. 95% of affected individuals carry pathogenic mutations in the *NPC1* gene (Patterson et al. 2012), which recently attracted attention because of its role in Ebola virus entry (Carette et al. 2011; Cote et al. 2011; White and Schornberg 2012). By contrast, the *NPC1* paralog, *NPC1L1* is not associated with the disease. On the contrary, inactivating mutations in *NPC1L1* reduce the risk of coronary heart disease (Myocardial Infarction Genetics Consortium et al. 2014). *NPC1* deletion in mice causes hearing loss (King et al. 2014), defects in retina (Yan et al. 2014), and deficiency in cerebellum development (Nusca et al. 2014), whereas *NPC1L1* deficiency protects ApoE-/- mice against atherosclerosis (Davis et al. 2007). Clearly, these two paralogs do not share identical functions and have different roles in health and disease.

Diagnosis of NP-C is challenging because of the heterogeneity in symptoms and clinical presentation (Vanier 2010). Until recently, the diagnostic standard was filipin staining of unesterified cholesterol in fibroblasts obtained by skin biopsy (Bornig and Geyer 1974; Vanier and Latour 2015). This test, however, is able to make a definitive diagnosis in only ~2/3 of cases. NP-C diagnostics has been significantly improved through the discovery of cholesterol oxidation products (“oxysterols”) that are elevated in the plasma of NP-C subjects (Porter et al. 2010). The plasma oxysterol assay detects >97% of cases with 100% sensitivity (Jiang et al. 2011). DNA sequencing offers another tool for NP-C diagnostics, but in practice detects only ~85% of NP-C cases due to the large number of private and non-coding sequence mutations (Stampfer et al. 2013). If a variant found in genetic testing has not been previously found to be disease causing, it is followed by risk estimation for pathogenicity. For novel missense mutations, *in silico* tools are indispensable to predict potential NP-C. However, each of the tools uses a different algorithm and some of them even use different data sets to evaluate the variant effect. For this reason, substantial inconsistencies between *in silico* tools are observed (Castellana and Mazza 2013). In case of a contradiction, deciding which software to trust in pathogenicity prediction remains a challenge. Researchers usually rely upon agreement between several tools, which has the effect of increasing specificity while decreasing the sensitivity (Wassif et al. 2015). Moreover, computational risk prediction tools that use conservation information do not discriminate between orthologous and paralogous proteins (Adzhubei et al. 2010), and, thus, include NPC1 paralogs, such as NPC1L1, in their analysis. Although including paralogs in risk estimating datasets is convenient (this eliminates computationally demanding and often non-trivial steps to separate orthologs and paralogs), such simplification confounds the function-specific signal.

NP-C disease caused by *NPC1* mutations is an ideal case study to understand the effects of paralogs in predicting disease causing mutations, because of a dramatic consequence of the duplication event that yielded *NPC1L1*. Moreover, many experimentally validated disease-causing mutations as well as alleles with high frequencies that are likely to be benign are known for this gene. We hypothesized that it is possible to predict the pathogenicity of single amino acid variants (SAVs) in NPC1 using only functionally equivalent human NPC1 (HsNPC1) orthologs.

In this study, we established the precise evolutionary history of the *NPC1* gene and identified evolutionary events that have likely affected its function. We used this information to build a computational approach, which showed improved accuracy in categorizing damaging and benign single amino acid substitutions in NPC1.

5.3 Results

Distinct Clusters of NPC1 Homologs Suggest Different Functions

NPC1 protein is predicted to have 13 transmembrane (TM) regions with 3 luminal domains. The crystal structure of the N-terminal domain has been solved with bound cholesterol, implicating this domain is involved in cholesterol binding and transport (Kwon et al. 2009). The pentahelical sterol-sensing domain, which resides between TM3 and TM8, likely responds to membrane cholesterol content and is required for cholesterol egress from the lysosome. There are 9 human genes, which share homology through their sterol-sensing domains and are identifiable in conventional sequence similarity searches initiated with NPC1: NPC1, NPC1-L1, PTCH1, PTCH2, PTCHD2, PTCHD3, PTCHD4, SCAB SREBF and DISP. These related proteins also share the “Patched” domain, which has a role in cholesterol-dependent processes. By contrast, domain architectures of these proteins show significant differences, where only NPC1 and NPC1-L1 contain the N-terminal cholesterol-binding domain **(Figure 5-1A)**. A phylogenetic tree constructed from the multiple sequence alignment of all Patched domain proteins shows distinct clades, where the NPC1-NPC1L1 clade is clearly separated from the rest of the Patched-containing sequences **(Figure 5-1B)**. These findings strongly suggest that other Patched-containing sequences should not be taken into account when examining function-specific characteristics of NPC1. In contrast, automated tools often include such functionally unrelated sequences in their datasets (see appendix).

Major events in NPC1 evolution

The *NPC1* gene is found in four of the five eukaryotic supergroups - unikonta, plants, chromalveolata and excavates – and is missing from Rhizeria. Phylogenetic analysis of NPC1 protein shows that the *NPC1* gene followed vertical evolution. Thus, it is likely that NPC1 was present in the last eukaryotic common ancestor (LECA). Multiple gene duplication events are observed in a range of taxonomic ranks from superorder to species level: among 397 species having *NPC1*, 195 (49%) have more than one copy (see appendix).

In the common ancestor of gnathostomata (jawed vertebrates), the *NPC1* gene was duplicated giving rise to the “NPC1-like” protein, which is present in most jawed vertebrates including humans (named NPC1L1). The NPC1L1 clade is greatly diverged from the root when compared to the gnathostomatan NPC1 clade **(Figure 5-2)**. NPC1 is present in each organism that has NPC1-L1; however, the opposite is not true. NPC1L1 is missing from some jawed vertebrate genomes. Moreover, the NPC1L1 clade has a longer average branch length from its root indicating a greater divergence **(Figure 5-2)**. The NPC1L1 divergence and dispensability strongly suggests that its function is different

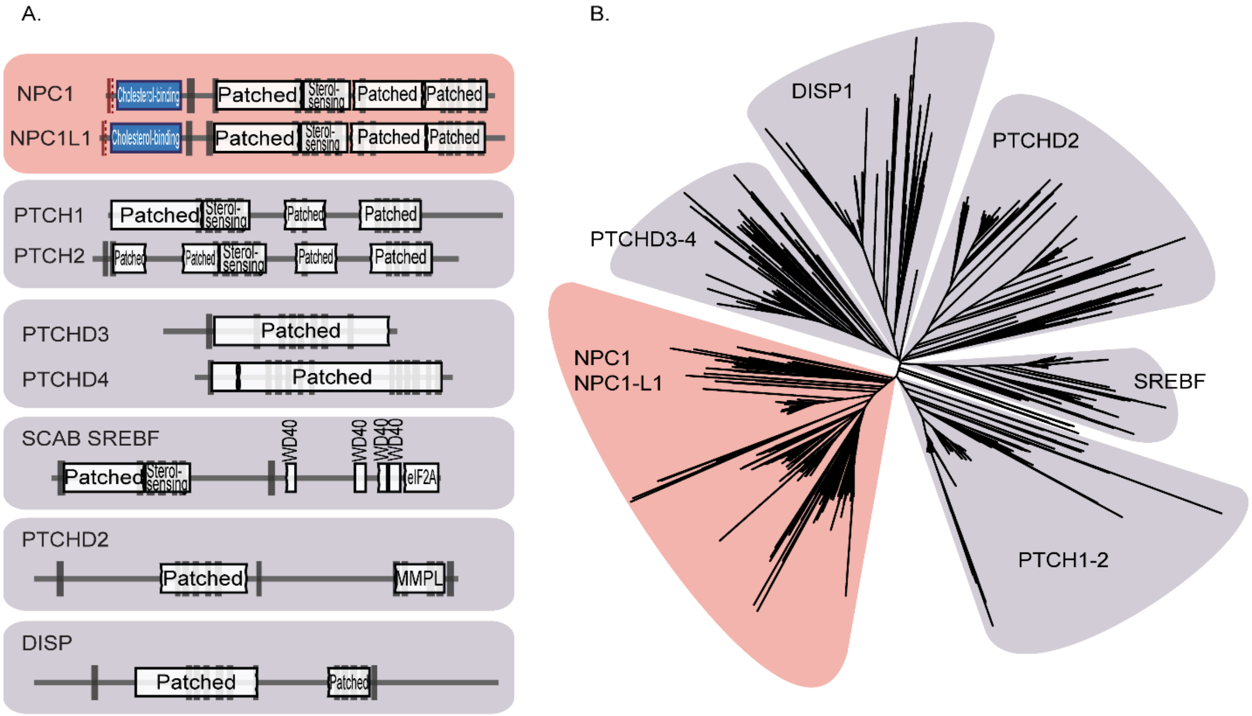


Figure 5-1 Relationships between Patched domain-containing proteins.

A) The domain architectures of human Patched domain-containing proteins were retrieved using the CDvist web server. Boxes with white background represent PFAM domains. Cholesterol-binding domains (in blue) were retrieved using a PDB database profile. Cholesterol-binding domain was found exclusively in NPC1 and NPC1L1. B) Some pairs such as PTCH1-PTCH2, NPC1-NPC1L1, PTCHD3-PTCHD4 have a relatively recent common ancestor, whereas the other proteins are related to each other more distantly, as they are represented as single clades on the phylogenetic tree. According to the phylogenetic tree the NPC1-NPC1L1 clade is clearly separated from other patched domain containing sequences.

\

from that of NPC1, which is further supported by the observation that no mutations in *NPC1-L1* have been associated with the Niemann-Pick C disease.

We observed another duplication in neoptera. As in gnathostomata, one of the two copies diverged from the original protein. Except for *Drosophila willistoni*, each neopteran genome containing the “diverged” copy also has the “original” version of the *NPC1* gene. However, the diverged copy is dispensable for some flies. In addition, because the diversified neopteran NPC1 shows higher within-clade divergence, it is likely to have gained a different function compared to the original protein, as seen in vertebrates (see appendix).

In fungi and amoebozoa, several duplications took place, but only at the species and genus level. So there was no major duplication event in these kingdoms.

In plants, there was a *NPC1* duplication in the common ancestor of flowering plants. More than one paralog is observed in *Pentapetalae*. However, the distances of two clades from the root are comparable **(Figure 5-2)**. Furthermore, some organisms have only one version of the gene from either clade, which suggests that one paralog is sufficient and neither copy is indispensable. Internal diversity of two clades were not significantly different from each other. Therefore, the clades may not have gained significantly different functions. For this reason, the *Homo sapiens* NPC1 (HsNPC1) orthology assignment cannot be precisely performed in plants.

Unikonts (metazoa, fungi and amoebozoa) and plants have the full-length NPC1 protein with an approximate length of 1300 amino acids and 13 TM regions, except for

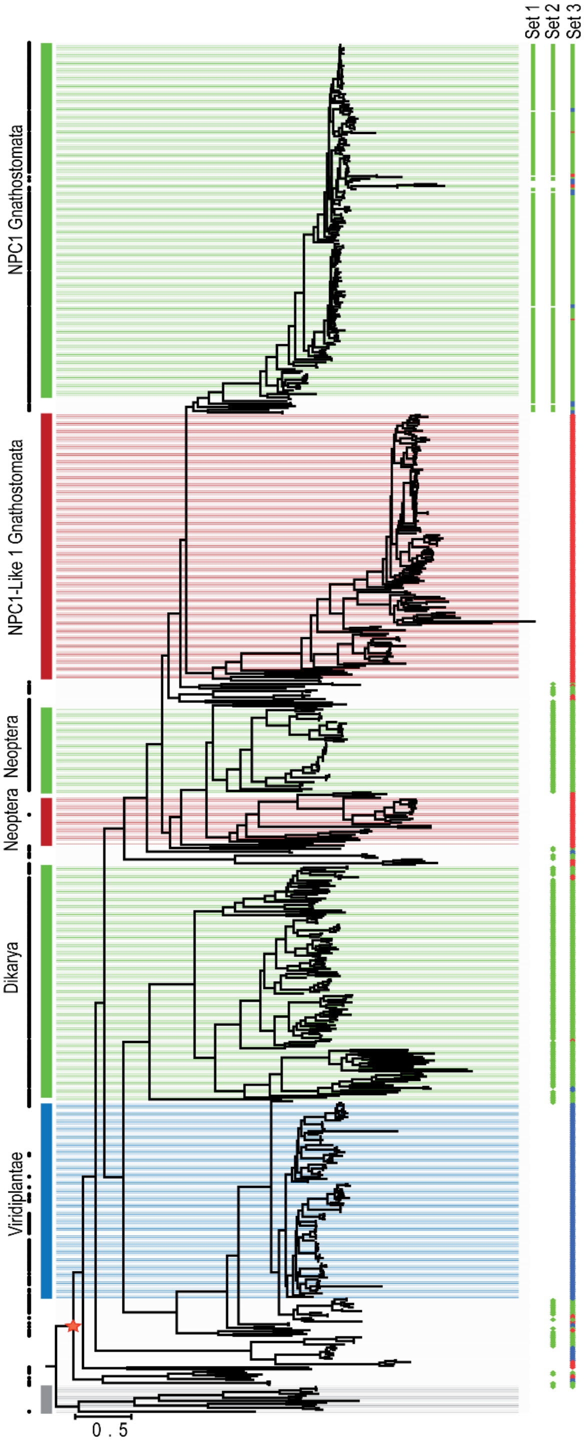


Figure 5-2 Maximum-likelihood phylogenetic tree of NPC1 proteins and described sets.

The star is placed at the root of full length NPC1. On the left side, the black markers represent the closest *NPC1* to the root for each organism. Green markers (Set 2) show the orthologs whereas red markers point to paralogs. Blue markers represent sequences which are ambiguous in terms of orthology. Gray-shaded clade contains short version of NPC1.

Dictyostelium, where two additional TM regions are inserted after TM-1 (see appendix). They all accommodate a lumenal N-terminal domain that binds to cholesterol. However, in *Naegleria gruberi* (excavate) and in most chromalveolates, the N-terminal cholesterol-binding domain is missing resulting in a shorter protein with 12 TM regions (see appendix). We found that all organisms that lack the NPC1 N-terminal domain, have a separate protein (~ 300 amino acids) encoded in their genomes, which is homologous (~30% identity, ~50% similarity) to the N-terminal domain of the full-length HsNPC1. Oomycetes have both “full” and “short” versions of NPC1. In the phylogenetic tree, these two versions are distinctly separated. Except for *Nannochloropsis gaditana* (which has an atypical NPC1 with no sterol-sensing domain), all organisms having the short version of NPC1 protein, also have the separate cholesterol binding protein. Moreover, the separate cholesterol binding protein is found exclusively in the organisms that have the short NPC1. The separate cholesterol binding protein is predicted to have a signal peptide at the N-terminus and a TM region at the C-terminus. Thus, concatenation of the separate cholesterol-binding protein and the short version of NPC1, substantially resembles HsNPC1. Exclusive coexistence of these two proteins suggests that they interact and function similarly to the full version of NPC1. The existence of both versions in oomycetes and the vertical evolutionary patterns suggest that both versions could have been present in LECA, where either fusion or dissociation could have occurred; then only one version was kept in all organisms, except for oomycetes, where both were kept.

In addition to major duplication events, in each kingdom, there were also species and/or genus level duplications. In such cases, we usually observe that in an organism, one copy evolves slowly to keep the original function, while the extra copies, which are not prone to purifying selective pressure, diverge faster. We used the distance measurements from the common ancestor node in the phylogenetic tree to determine the “incomparably least diverged” (slowest evolving) gene, which in turn enables us to find the functional orthologs. However, in some cases orthology assignment was inconclusive due to comparable divergence behaviors.

NPC1 was lost in many parasites including whole clades, such as microsporidia (fungi) and apicomplexa (chromalveolata). Except for *Naegleria gruberi*, all species sequenced in the excavate supergroup are parasitic (*Trypanosomatidae* family, *Trichomonas vaginalis* and *Giardia intestinalis*) and contain no *NPC1* in their genomes.

Defining HsNPC1 Functional Orthologs

Products of orthologous genes are very likely to perform the same function. Therefore, distinguishing HsNPC1 orthologs from other homologous proteins is critical in order to identify potentially pathogenic variants specifically affecting HsNPC1 function. Detailed analysis of the phylogenetic tree of all NPC1 homologs guided HsNPC1 orthology assignment. The clades retaining the original NPC1 function were determined based on the agreement of three lines of evidence. First, we compared the distances of duplicated clades to the full-length NPC1 root **(Figure 5-2)** to identify which one is less diverged. Second, we compared the organism content of the clades. If a clade is subset of another, then the superset clade was considered the “original” one representing HsNPC1 orthologs. Finally, diversity within the clades was assessed: the less diverged clade is more likely to be ancestral (see appendix). When all three criteria agree, HsNPC1 orthologs can be identified with confidence. However, in some cases, the sequence divergence information was inconclusive. In those cases, none of the clades was a subset of another. Moreover, the diversity within the clades was comparable. Consequently, these sequences were classified as "ambiguous" and they were not included in the set of HsNPC1 orthologs.

Evaluating Missense Mutations in HsNPC1: the Scoring Algorithm

Our master multiple sequence alignment (MSA) included all homologs, such as paralogs and short version duplicates. We divided the master MSA into three sets grouped by the orthology relationships (see Figure 2 for details). The phylogenetic clade containing HsNPC1 after the most recent major evolutionary event, which is the birth of NPC1L1 in gnathostomata, was considered as the core alignment. Not surprisingly, this alignment set had a high level of sequence conservation. We refer to this alignment as "Set 1". Set 1 is given the highest importance in the evaluation algorithm. Set 2 includes Set 1 and also other sequences which have unambiguous one-to-one orthology with HsNPC1. Finally, Set 3 contained all HsNPC1 homologs, including paralogs and "ambiguous" orthologs, except for the short versions of NPC1.

In order to predict the effect of missense mutations on HsNPC1 function, we propose an algorithm (SAVER: Single Amino Acid Variant Evaluator) that provides binary output from the MSA analysis of Sets 1 and 2 **(Figure 5-3)**. In the scoring part, Set 1 is given the highest weight, because it contains HsNPC1 and its orthologs that evolved after the most recent duplication (MRD). The birth of many Mendelian diseases correlates with the time of MRDs (Dickerson and Robertson 2012). However, using only Set 1, which is limited in our case to bilaterian genomes, would not be sufficient in collecting the entire ancestral information. For this reason, Set 2 was used to compensate for the lack of evolutionary depth in Set 1. Because Set 2 was carefully constructed from sequences that are likely to conserve the ancestral function of NPC1, the amount of false signal it introduces is limited. Furthermore, the possibility of false signals in Set 2 was addressed by lowering its priority. Because sufficient evolutionary depth was reached with set 2, specificity was not affected drastically by excluding Set 3-only sequences.

Sequencing and aligning errors are key factors causing misinterpretation. For example, a pathogenic variant can be categorized as benign, when the corresponding position in MSA appears variable due to several misaligned sequences. For this reason, working with the cleanest possible data set, a nearly perfect alignment and well-constructed phylogenetic trees is critical in assessing the mutations. *Ab initio* elimination of sequences that have misaligned regions is not an optimum solution, because these sequences may also contain well-aligned regions carrying important information. In our approach we apply positional masking of misaligned regions, so that well-aligned positions in these sequences are taken into account. Another challenge in eukaryotic sequence comparison is dealing with isoforms, which are different protein products of the same gene due to alternative splicing. The isoforms can redundantly dominate the signal and cause artificial conserved positions. Moreover, on the borders of alternative splicing, the unrelated sequences of isoforms can be aligned together. We resolve this issue by choosing a representative isoform for each gene. Selection of a representative isoform depends on the queried position, in order to rule out the errors that alternative-splicing prediction can cause.

For a single amino acid substitution from AA_0 to AA_1, scoring algorithms usually use the abundance of the AA_1 in MSA. However, instead of counting the number of sequences

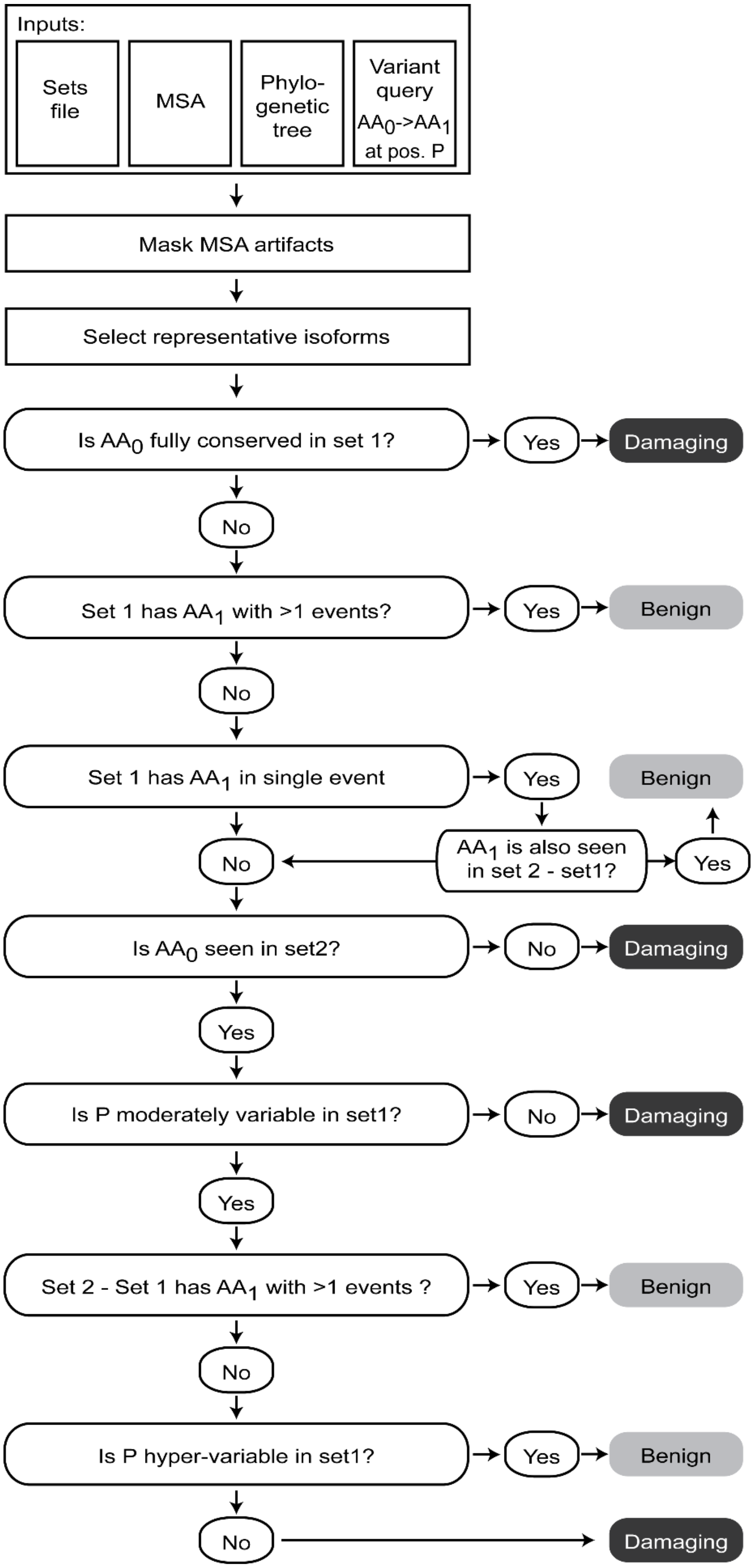

Inputs:
Sets file
MSA
Phylo-genetic tree
Variant query AA_0->AA_1 at pos. P
Mask MSA artifacts
Select representative isoforms
Is AA_0 fully conserved in set 1?
Yes
Damaging
No
Set 1 has AA_1 with >1 events?
Yes
Benign
No
Set 1 has AA_1 in single event
Yes
Benign
AA_1 is also seen in set 2 - set1?
Yes
No
Is AA_0 seen in set2?
No
Damaging
Yes
Is P moderately variable in set1?
No
Damaging
Yes
Set 2 - Set 1 has AA_1 with >1 events ?
Yes
Benign
No
Is P hyper-variable in set1?
Yes
Benign
No
Damaging

Figure 5-3 SAVER algorithm workflow.

P: position of the substitution; AA_0: original amino acid; AA_1: replacing amino acid.

with substitutions, we propose to count how many times a given replacement has occurred independently, so a single evolutionary event would not be counted multiple times. Distinguishing between single and multiple independent substitutions is critical, because multiple independent substitutions, occurring in different clades, suggest that a position tolerates mutations, whereas a single substitution compensated by a suppressor mutation can be in a potentially “irreplaceable” position. It is important to stress that such information can only be obtained from well-edited multiple sequence alignments and well-built phylogenetic trees that require substantial manual work.

Improved success in distinguishing between damaging and benign SAVs.

We scanned literature to retrieve known NPC1 variants. Only single amino acid substitutions were taken into account. Only biochemically validated NP-C causing mutations were considered as “damaging” variants. Recently published frequencies of HsNPC1 variants from several exome sequencing data sets (Wassif et al. 2015) were used to define the benign mutation data set. We selected the common variants that have never been shown as pathogenic in any study, and that have frequency greater than 0.028%, which is the frequency of the most commonly reported pathogenic variant, I1061T. Our compiled control set contained 166 damaging and 21 benign SAVs (see appendix).

We tested our approach in comparison with leading automated tools: PolyPhen-2, SIFT and PROVEAN (Ng and Henikoff 2003; Adzhubei et al. 2010; Choi and Chan 2015). The results indicate that our approach outperforms other tools (i) in terms of sensitivity (~10% improvement), while causing a relatively low cost in specificity and (ii) in terms of the overall quality, as measured by the Matthews correlation coefficient **(Table 5-1)**. The drastic improvement in sensitivity can be explained by the fact that our method eliminates the false evolutionary signals introduced by functionally diverged sequences that are included in the analysis by other tools (see appendix).

We also applied our method to all theoretical amino acid substitutions in NPC1. 24282 (1278 positions in NPC1 sequence X 19 amino acid substitutions) theoretical SAVs were evaluated by our approach in comparison with the well optimized automated methods described above (see attachment). Ultimately, our method predicts 81% of the variants as damaging, while PolyPhen-2, PROVEAN and SIFT predict 60%, 70% and 66% as damaging, respectively. Because we suspected that our approach over predicts damaging variants, we adjusted the cutoffs of other tools to fix the damaging rate at 81%. After the adjustment, the performance of two methods (PolyPhen-2 and PROVEAN) was improved; however, none of them reached the quality of our approach, as measured by Matthews Correlation Coefficient value. Comparison between receiver operating characteristics of the tools and our “sensitivity - false positive rate” datum, shows a clear distinction of our result from the general trend of the others (see appendix).

An example of how inclusion of paralogous sequences negatively affects the prediction is shown in Figure 4. Known pathogenic mutations, N968S, G986S, G993A and M995R (see appendix) are predicted as benign by all three automated tools, because the same substitutions are found in NPC1L1 paralogs that are included in their MSA sets **(Figure 5-4)**. Figure 5-5 shows topology of the human NPC1 where the positions are colored based on the numbers of allowed amino acids at that position by our approach. This risk map provides clues about the functionally critical regions of HsNPC1 (see appendix) and the full list of potentially damaging and benign substitutions in this protein is provided as

Table 5-1 Performance comparison of tools predicting the effect of NPC1 missense mutations.

	Damaging (166)		**Benign (21)**		**Sensitivity**	**Specificity**	**False Discovery Rate (FDR)**	**Accuracy**	**F1**	**MCC**
	TP	FN	TN	FP						
SAVER	157	9	14	7	0.95	0.67	0.04	0.91	0.95	0.59
PP2	139	27	15	6	0.84	0.71	0.04	0.82	0.89	0.42
PROVEAN	141	25	15	6	0.85	0.71	0.04	0.83	0.90	0.43
SIFT	135	31	18	3	0.81	0.86	0.02	0.82	0.89	0.48
PP2$_{\textbf{adj}}$	159	7	12	9	0.96	0.57	0.05	0.91	0.95	0.55
PROVEAN$_{\textbf{adj}}$	153	13	12	9	0.92	0.57	0.06	0.88	0.93	0.46
SIFT$_{\textbf{adj}}$	150	16	11	10	0.90	0.52	0.06	0.86	0.92	0.38

The cutoffs distinguishing between "damaging" and "benign" variants, are changed in the methods which have subscripted with "adj" abbreviation based on the output of the SAVER computation. SAVER and other "adjusted" tools yield 81% damaging rate in all theoretical amino acid substitutions on HsNPC1. TP: True positive; FN: False negative; TN: True negative; FP: False positive.

attachment. We have built a web-based application for querying single amino acid variants in NPC1, which can serve as a reference for clinicians when describing novel NP-C causing mutations. It is freely available at http://genomics.utk.edu/saver/npc1.html.

5.4 Discussion

In this work, we showed that it is possible to get closer to the desired level in predicting the effects of missense mutations by carefully analyzing the evolutionary history of a gene. A clear improvement is accomplished by taking into consideration only function-specific orthologous protein sequences. Remote homologs and paralogs that are likely to be functionally diverged should be removed from the analysis. In selecting functional counterparts, specific criteria based on a thorough phylogenetic analysis must be used.

The proposed approach heavily depends on manual work (constructing high-quality datasets, alignments, trees and defining orthologs and paralogs) as well as reasoning, which depends on the output of a particular computational step. Thus, for now, this approach cannot be fully automated and will not replace any of the available automated tools. However, revealing common trends and problems in identifying functional orthologs and testing this approach on other well-defined monogenic Mendelian diseases, should lead to the development of the next generation of predictive automated methods directly applicable in clinical practices.

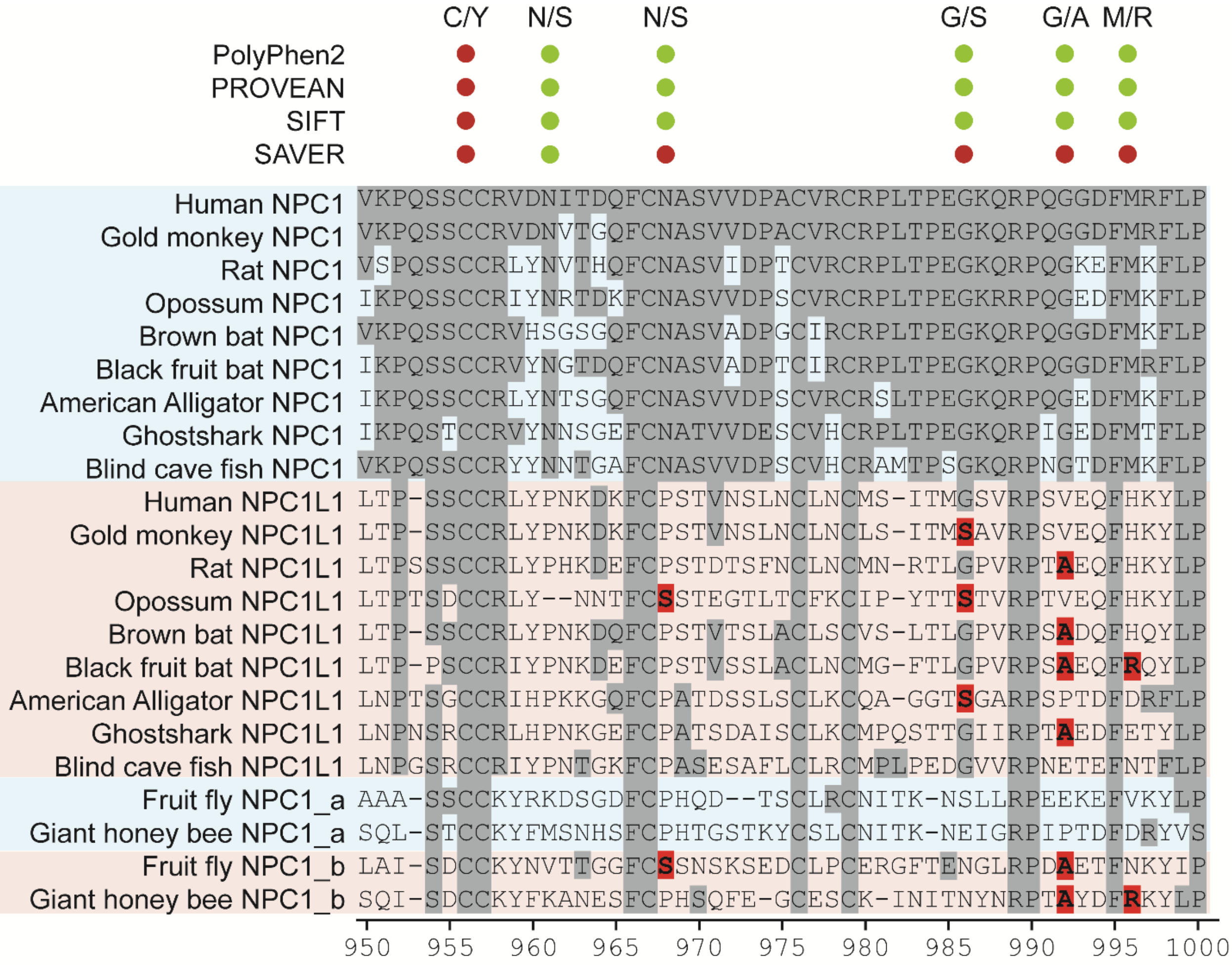

C/Y
N/S
N/S
G/S
G/A
M/R
PolyPhen2
PROVEAN
SIFT
SAVER
Human NPC1 VKPQSSCCRVDNITDQFCNASVVDPACVRCRPLTPEGKQRPQGGDFMRFLP
Gold monkey NPC1 VKPQSSCCRVDNVTGQFCNASVVDPACVRCRPLTPEGKQRPQGGDFMRFLP
Rat NPC1 VSPQSSCCRLYNVTHQFCNASVIDPTCVRCRPLTPEGKQRPQGKEFMKFLP
Opossum NPC1 IKPQSSCCRIYNRTDKFCNASVVDPSCVRCRPLTPEGKRRPQGEDFMKFLP
Brown bat NPC1 VKPQSSCCRVHSGSGQFCNASVADPGCIRCRPLTPEGKQRPQGGDFMKFLP
Black fruit bat NPC1 IKPQSSCCRVYNGTDQFCNASVADPTCIRCRPLTPEGKQRPQGGDFMRFLP
American Alligator NPC1 IKPQSSCCRLYNTSGQFCNASVVDPSCVRCRSLTPEGKQRPQGEDFMKFLP
Ghostshark NPC1 IKPQSTCCRVYNNSGEFCNATVVDESCVHCRPLTPEGKQRPIGEDFMTFLP
Blind cave fish NPC1 VKPQSSCCRYYNNTGAFCNASVVDPSCVHCRAMTPSGKQRPNGTDFMKFLP
Human NPC1L1 LTP-SSCCRLYPNKDKFCPSTVNSLNCLNCMS-ITMGSVRPSVEQFHKYLP
Gold monkey NPC1L1 LTP-SSCCRLYPNKDKFCPSTVNSLNCLNCLS-ITMSAVRPSVEQFHKYLP
Rat NPC1L1 LTPSSSCCRLYPHKDEFCPSTDTSFNCLNCMN-RTLGPVRPTAEQFHKYLP
Opossum NPC1L1 LTPTSDCCRLY--NNTFCSSTEGTLTCFKCIP-YTTSTVRPTVEQFHKYLP
Brown bat NPC1L1 LTP-SSCCRLYPNKDQFCPSTVTSLACLSCVS-LTLGPVRPSADQFHQYLP
Black fruit bat NPC1L1 LTP-PSCCRIYPNKDEFCPSTVSSLACLNCMG-FTLGPVRPSAEQFRQYLP
American Alligator NPC1L1 LNPTSGCCRIHPKKGQFCPATDSSLSCLKCQA-GGTSGARPSPTDFDRFLP
Ghostshark NPC1L1 LNPNSRCCRLHPNKGEFCPATSDAISCLKCMPQSTTGIIRPTAEDFETYLP
Blind cave fish NPC1L1 LNPGSRCCRIYPNTGKFCPASESAFLCLRCMPLPEDGVVRPNETEFNTFLP
Fruit fly NPC1_a AAA-SSCCKYRKDSGDFCPHQD--TSCLRCNITK-NSLLRPEEKEFVKYLP
Giant honey bee NPC1_a SQL-STCCKYFMSNHSFCPHTGSTKYCSLCNITK-NEIGRPIPTDFDRYVS
Fruit fly NPC1_b LAI-SDCCKYNVTTGGFCSSNSKSEDCLPCERGFTENGLRPDAETFNKYIP
Giant honey bee NPC1_b SQI-SDCCKYFKANESFCPHSQFE-GCESCK-INITNYNRPTAYDFRKYLP
950
955
960
965
970
975
980
985
990
995
1000

Figure 5-4 An alignment window illustrating false effects of paralogs in predicting damaging mutations.

Blue shaded sequences are HsNPC1 orthologs and the rest are paralogs. For each tool, a red marker represents "predicted as damaging", whereas green marker stands for "predicted as benign". Residues highlighted in red are the potential causes of predicting pathogenic variants as benign.

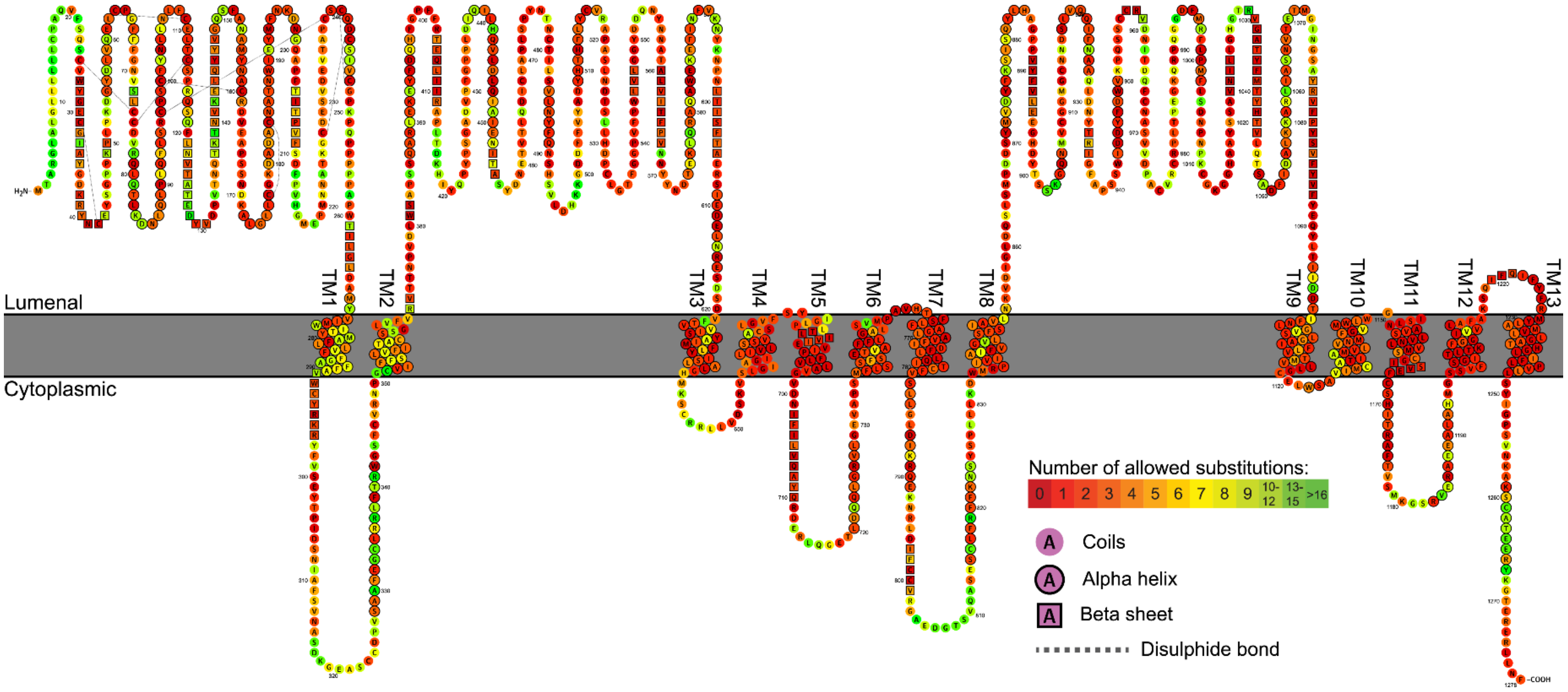


Figure 5-5 NPC1 missense mutation risk map.

Color scale from red to green ascendingly shows the number of amino acid substitutions that are predicted to be benign. Secondary structure information was retrieved from a 3D structure for the cholesterol-binding domain (PDB ID: 3GKH) and predicted for the rest of the protein. Squared residues represent beta-sheet, while circles with outline stand for alpha helices. Disulphide bonds are represented (dashed lines) only for the cholesterol-binding domain.

5.5 Materials and Methods

Databases, multiple sequence alignments and phylogenetic trees

Human NPC1 protein (NM_000271.4) was queried through blastp (Altschul et al. 1990) against the human genome to reveal the related sequences. Each hit was blasted individually against the RefSeq database (Pruitt et al. 2012). For each job the full sequences were compiled and aligned using MAFFT default algorithm (Katoh and Standley 2013). Neighbor joining tree was built with the phylip package (Retief 2000). From the tree, the NPC1-homologs clade was isolated. With the retrieved homologs, MAFFT version v7.154b E-INS-i algorithm was used to realign the full-length sequences. A maximum-likelihood tree was constructed using the PhyML software version 20140929 (Guindon et al. 2009), with JTT substitution model (Jones et al. 1992) and the remaining parameters as default. The outgroups that were not considered to be NPC1 homologs based on Refseq annotations and domain architectures were discarded from the multiple sequence alignment, NPC1 homologs were realigned and the final phylogenetic tree was built using the previously described approach.

Taxonomic distribution

After obtaining the final set of sequences, gene IDs were assigned to protein sequences using NCBI Entrez (Gibney and Baxevanis 2011). The gene counts were visualized on a taxonomically classified sunburst tree. Taxonomical ranks were taken from the NCBI taxonomy database (Federhen 2012). Organisms were selected based on two criteria: (i) the availability of their NPC1 in the RefSeq database and (ii) the availability of their genome in the NCBI genome database. The sunburst visualization was performed with a custom built tool.

Orthology assignment

Orthologs and paralogs were distinguished using the maximum likelihood phylogenetic tree. In case of major duplication events, a consistently more divergent duplicated clade was categorized as paralogs that are less likely to retain the original NPC1 function. The reference point for evolutionary distance was determined as the full-length NPC1 node. In the cases where no divergence consistency between clades was observed (e.g. not all species in clade A were more diverged than those in clade B, or incomplete species set in both clades), the orthology assignment was deemed inconclusive. In such cases, we considered both clades as paralogs that have a potential to gain a modified function. For the species-level duplications, the sequence, which was significantly diverged from the closest node of NPC1 orthologs, was categorized as paralogous.

Scoring the effect of single amino acid variants

PubMed 1997-2014 database was manually searched to identify relevant studies and case series. The search key words used were: (i) “Niemann-Pick type C”, (ii) “NPC1”, (iii) “NPC1 mutations”. No other search restrictions were applied and all related reference articles were retrieved and reviewed. The initial search resulted in 312 papers. General review articles on Niemann-Pick disease type C pathogenesis, course and outcomes, basic clinical case reports lacking genetic testing and experimental findings not connected with clinical data were excluded. As a result, we identified 56 articles referencing a total of 572 mutations in the *NPC1* gene (including repetitive reports). After refining this list by excluding repetitive reports, insertion/deletion, frameshift and nonsense mutations and benign SNPs, the final list of most likely pathogenic SAVs was comprised of 166 variants that were referred to as “damaging” variants in this study. In order to retrieve the set of “benign” mutations, we used frequencies in human populations reported by Wassif et al. (Wassif et al. 2015). The variants found in humans with higher frequency than the most common deleterious variant, I1063T, were categorized as benign. However, we removed N222S, N961S, S1200G and A521S from this list, due to the reports suggesting that they might be damaging.

In our algorithm, the “moderately variable” category was defined as a position having more than 5 different substitutions in a given set. Position was categorized as “hyper-variable” if there were more than 9 different substitutions.

Statistical analyses

The performance of the algorithm is described by the following parameters: sensitivity, specificity, false discovery rate, accuracy, F1 score and Matthews Correlation Coefficient (MCC). In the equations given below, TP, TN, FP and FN refer to the number of true positives, true negatives, false positives and false negatives, respectively.

$$Sensitivity = \frac{\mathrm{TP}}{\mathrm{TP} + \ \mathrm{FN}}$$

$$Specificity = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

$$False\ Discovery\ Rate = 1 - \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$F1 = 1 - \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

$$MCC = \frac{\mathrm{TP\ X\ TN} - \mathrm{FP\ X\ FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(TP + FN)(TN + FN)(TN + FN)}}$$

Domain architecture prediction and risk map generation

We used CDvist webserver with HMMER3 against Pfam 27.0 and HHsearch against PDB options respectively (Finn et al. 2014; Adebali et al. 2015). PDB HHsearch probability cutoff was adjusted to 98%. Transmembrane regions and signal peptides were predicted using Phobius (Kall et al. 2007).

We implemented the SAVER algorithm in a python3 script and ran it on all theoretical human NPC1 SAVs. For each position, we counted the allowed (benign) amino acids. The range was between 0 (no substitution allowed) and 19 (any substitution allowed). For secondary structure information, X-ray crystal structure (PDB ID: 3GKH) was used for N-terminal domain and Psipred prediction was used for the rest. Protter web application was used to generate the NPC1 membrane topology figure using default parameters for transmembrane region and signal-peptide predictions. Disulphide bond information for N-terminal domain was also collected from the available structure.

5.6 Acknowledgments

This work was supported in part by the National Institute of Health (grant number GM072295 to I.B.Z.).

5.7 References

Adebali O, Ortega DR, Zhulin IB. 2015. CDvist: a webserver for identification and visualization of conserved domains in protein sequences. *Bioinformatics* **31**: 1475-1477.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods* **7**: 248-249.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403-410.

Bornig H, Geyer G. 1974. Staining of cholesterol with the fluorescent antibiotic "filipin". *Acta Histochem* **50**: 110-115.

Carette JE, Raaben M, Wong AC, Herbert AS, Obernosterer G, Mulherkar N, Kuehne AI, Kranzusch PJ, Griffin AM, Ruthel G et al. 2011. Ebola virus entry requires the cholesterol transporter Niemann-Pick C1. *Nature* **477**: 340-343.

Castellana S, Mazza T. 2013. Congruency in the prediction of pathogenic missense mutations: state-of-the-art web-based tools. *Brief Bioinform* **14**: 448-459.

Chang F, Li MM. 2013. Clinical application of amplicon-based next-generation sequencing in cancer. *Cancer Genet* **206**: 413-419.

Choi Y, Chan AP. 2015. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* doi:10.1093/bioinformatics/btv195.

Cote M, Misasi J, Ren T, Bruchez A, Lee K, Filone CM, Hensley L, Li Q, Ory D, Chandran K et al. 2011. Small molecule inhibitors reveal Niemann-Pick C1 is essential for Ebola virus infection. *Nature* **477**: 344-348.

Davis HR, Jr., Hoos LM, Tetzloff G, Maguire M, Zhu LJ, Graziano MP, Altmann SW. 2007. Deficiency of Niemann-Pick C1 Like 1 prevents atherosclerosis in ApoE-/- mice. *Arterioscler Thromb Vasc Biol* **27**: 841-849.

Dickerson JE, Robertson DL. 2012. On the origins of Mendelian disease genes in man: the impact of gene duplication. *Mol Biol Evol* **29**: 61-69.

Federhen S. 2012. The NCBI Taxonomy database. *Nucleic acids research* **40**: D136-143.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J et al. 2014. Pfam: the protein families database. *Nucleic acids research* **42**: D222-230.

Gibney G, Baxevanis AD. 2011. Searching NCBI Databases Using Entrez. *Curr Protoc Hum Genet* **Chapter 6**: Unit6 10.

Guindon S, Delsuc F, Dufayard JF, Gascuel O. 2009. Estimating maximum likelihood phylogenies with PhyML. *Methods Mol Biol* **537**: 113-137.

Jahnova H, Dvorakova L, Vlaskova H, Hulkova H, Poupetova H, Hrebicek M, Jesina P. 2014. Observational, retrospective study of a large cohort of patients with Niemann-Pick disease type C in the Czech Republic: a surprisingly stable diagnostic rate spanning almost 40 years. *Orphanet J Rare Dis* **9**: 140.

Jiang X, Sidhu R, Porter FD, Yanjanin NM, Speak AO, te Vruchte DT, Platt FM, Fujiwara H, Scherrer DE, Zhang J et al. 2011. A sensitive and specific LC-MS/MS method for rapid diagnosis of Niemann-Pick C1 disease from human plasma. *J Lipid Res* **52**: 1435-1445.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**: 275-282.

Jordan DM, Ramensky VE, Sunyaev SR. 2010. Human allelic variation: perspective from protein function, structure, and evolution. *Current opinion in structural biology* **20**: 342-350.

Kall L, Krogh A, Sonnhammer EL. 2007. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res* **35**: W429-432.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772-780.

Katsanis SH, Katsanis N. 2013. Molecular genetic testing and the future of clinical genomics. *Nature reviews Genetics* **14**: 415-426.

King KA, Gordon-Salant S, Pawlowski KS, Taylor AM, Griffith AJ, Houser A, Kurima K, Wassif CA, Wright CG, Porter FD et al. 2014. Hearing loss is an early consequence of Npc1 gene deletion in the mouse model of Niemann-Pick disease, type C. *J Assoc Res Otolaryngol* **15**: 529-541.

Kwon HJ, Abi-Mosleh L, Wang ML, Deisenhofer J, Goldstein JL, Brown MS, Infante RE. 2009. Structure of N-terminal domain of NPC1 reveals distinct subdomains for binding and transfer of cholesterol. *Cell* **137**: 1213-1224.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151-1155.

Myocardial Infarction Genetics Consortium I, Stitziel NO, Won HH, Morrison AC, Peloso GM, Do R, Lange LA, Fontanillas P, Gupta N, Duga S et al. 2014. Inactivating mutations in NPC1L1 and protection from coronary heart disease. *N Engl J Med* **371**: 2072-2082.

Ng PC, Henikoff S. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**: 3812-3814.

Ng PC, Henikoff S. 2006. Predicting the effects of amino acid substitutions on protein function. *Annual review of genomics and human genetics* **7**: 61-80.

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA et al. 2010. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**: 30-35.

Nusca S, Canterini S, Palladino G, Bruno F, Mangia F, Erickson RP, Fiorenza MT. 2014. A marked paucity of granule cells in the developing cerebellum of the Npc1(-/-) mouse is corrected by a single injection of hydroxypropyl-beta-cyclodextrin. *Neurobiol Dis* **70**: 117-126.

Oliver GR, Hart SN, Klee EW. 2015. Bioinformatics for clinical next generation sequencing. *Clin Chem* **61**: 124-135.

Patterson MC, Hendriksz CJ, Walterfang M, Sedel F, Vanier MT, Wijburg F, Group N-CGW. 2012. Recommendations for the diagnosis and management of Niemann-Pick disease type C: an update. *Mol Genet Metab* **106**: 330-344.

Porter FD, Scherrer DE, Lanier MH, Langmade SJ, Molugu V, Gale SE, Olzeski D, Sidhu R, Dietzen DJ, Fu R et al. 2010. Cholesterol oxidation products are sensitive and specific blood-based biomarkers for Niemann-Pick C1 disease. *Sci Transl Med* **2**: 56ra81.

Pruitt KD, Tatusova T, Brown GR, Maglott DR. 2012. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* **40**: D130-135.

Retief JD. 2000. Phylogenetic analysis using PHYLIP. *Methods Mol Biol* **132**: 243-258.

Sleat DE, Wiseman JA, El-Banna M, Price SM, Verot L, Shen MM, Tint GS, Vanier MT, Walkley SU, Lobel P. 2004. Genetic evidence for nonredundant functional cooperativity between NPC1 and NPC2 in lipid transport. *Proc Natl Acad Sci U S A* **101**: 5886-5891.

Stampfer M, Theiss S, Amraoui Y, Jiang X, Keller S, Ory DS, Mengel E, Fischer C, Runz H. 2013. Niemann-Pick disease type C clinical database: cognitive and coordination deficits are early disease indicators. *Orphanet J Rare Dis* **8**: 35.

Sunyaev SR. 2012. Inferring causality and functional significance of human coding DNA variants. *Hum Mol Genet* **21**: R10-17.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**: 64-69.

Vanier MT. 2010. Niemann-Pick disease type C. *Orphanet J Rare Dis* **5**: 16.

Vanier MT. 2015. Complex lipid trafficking in Niemann-Pick disease type C. *J Inherit Metab Dis* **38**: 187-199.

Vanier MT, Latour P. 2015. Laboratory diagnosis of Niemann-Pick disease type C: the filipin staining test. *Methods Cell Biol* **126**: 357-375.

Wang Z, Moult J. 2001. SNPs, protein structure, and disease. *Hum Mutat* **17**: 263-270.

Wassif CA, Cross JL, Iben J, Sanchez-Pulido L, Cougnoux A, Platt FM, Ory DS, Ponting CP, Bailey-Wilson JE, Biesecker LG et al. 2015. High incidence of unrecognized visceral/neurological late-onset Niemann-Pick disease, type C1, predicted by analysis of massively parallel sequencing data sets. *Genet Med* doi:10.1038/gim.2015.25.

White JM, Schornberg KL. 2012. A new player in the puzzle of filovirus entry. *Nature reviews Microbiology* **10**: 317-322.

Yan X, Ma L, Hovakimyan M, Lukas J, Wree A, Frank M, Guthoff R, Rolfs A, Witt M, Luo J. 2014. Defects in the retina of Niemann-pick type C 1 mutant mice. *BMC Neurosci* **15**: 126.

5.8 Appendix

Interpretation of the risk map

Each residue in a protein experiences a different selection pressure due to their molecular function. Some residues are replaceable, and some cannot be altered. Additionally, there are positions allowing replacements by only certain amino acids. Though interchangeable amino acids usually share at least one physiochemical characteristic, this is not always the case.

According to the NPC1 risk map **(Figure 5-5)**, first two TM regions are not conserved as well as other TM regions. Thus, these two regions may not be involved in sterol transport process. Instead, they could have evolved as a simple connector to join lumenal cholesterol-binding domain with the rest of the protein. On the other hand, the rest of the TM regions show moderate to high level of conservation. Particularly, TM5, TM11, TM12 and TM13 may play critical roles in the transport process. We also see heterogeneous high level of conservation in the lumenal domains. Predicted helices and beta sheets generally correlate with the conservation pattern. However, some predicted unstructured regions are also highly conserved. Specifically, most cysteine residues in the lumenal domains are invariable, likely because of their specific structural role of building disulfide bridges. The cytoplasmic regions between TM5 and TM6, and TM11 and TM12 are predicted to form secondary structure elements and also are well conserved, which indicates that these regions may play a distinct functional role.

The risk map built based on set 1 and set 2 shows evolutionary patterns that we already deduced from the MSA of all NPC1 homologs. For instance, TM1 and TM2 were considered to function in linking cholesterol-binding domain with the rest, based on the analysis of domain architectures of NPC1 homologs in chromalveolata and *Naegleria*. Even these sequences are removed from the analysis, we can still observe a relaxed selection pressure on these regions, which also indicates that they are not functionally critical. Taken together these observations suggest that the evolutionary depth of the dataset, which includes only functional orthologs of NPC1, is sufficient to infer tolerant and intolerant substitutions.

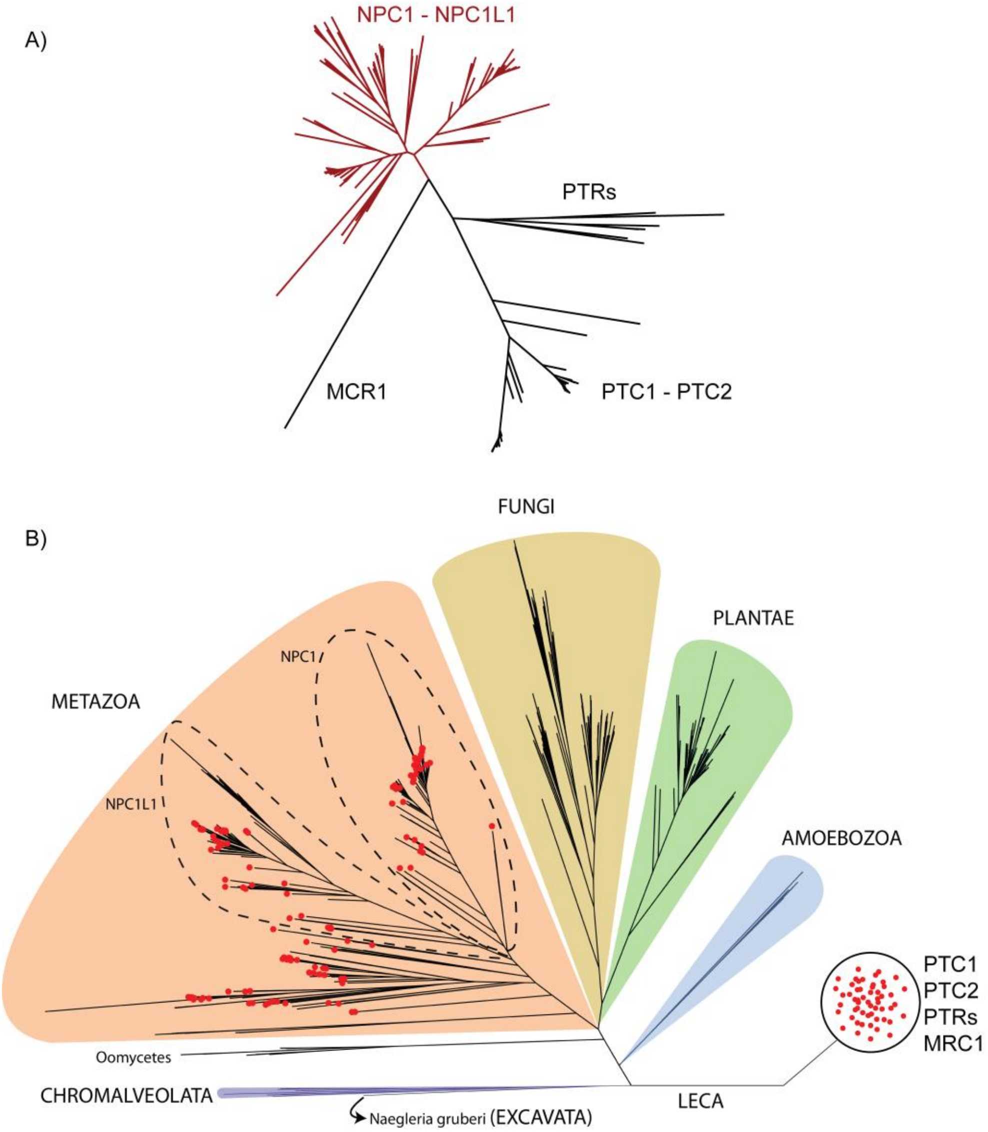


Figure 5-6 Common problems in automated prediction of functional effects of amino acid substitutions in proteins.

Close homologs are often missing from automatically constructed datasets, whereas functionally unrelated remote homologs might be present. A) The maximum likelihood phylogenetic tree of sequences from the alignment we generated based on raw sequences retrieved (February 15th, 2015) from PolyPhen-2 (PP2) human NPC1 (Uniprot ID: O15118) query. B) Maximum-likelihood phylogeny of NPC1 homologs from the core alignment compiled in this study. Red markers identify sequences that are present in the PP2 alignment. The unmarked sequences are missing from the PP2 alignment. PTC1-2: Patched protein homolog 1-2, PTRs: Unclassified patched receptor like proteins, MRC1: Mannose receptor, C type 1.

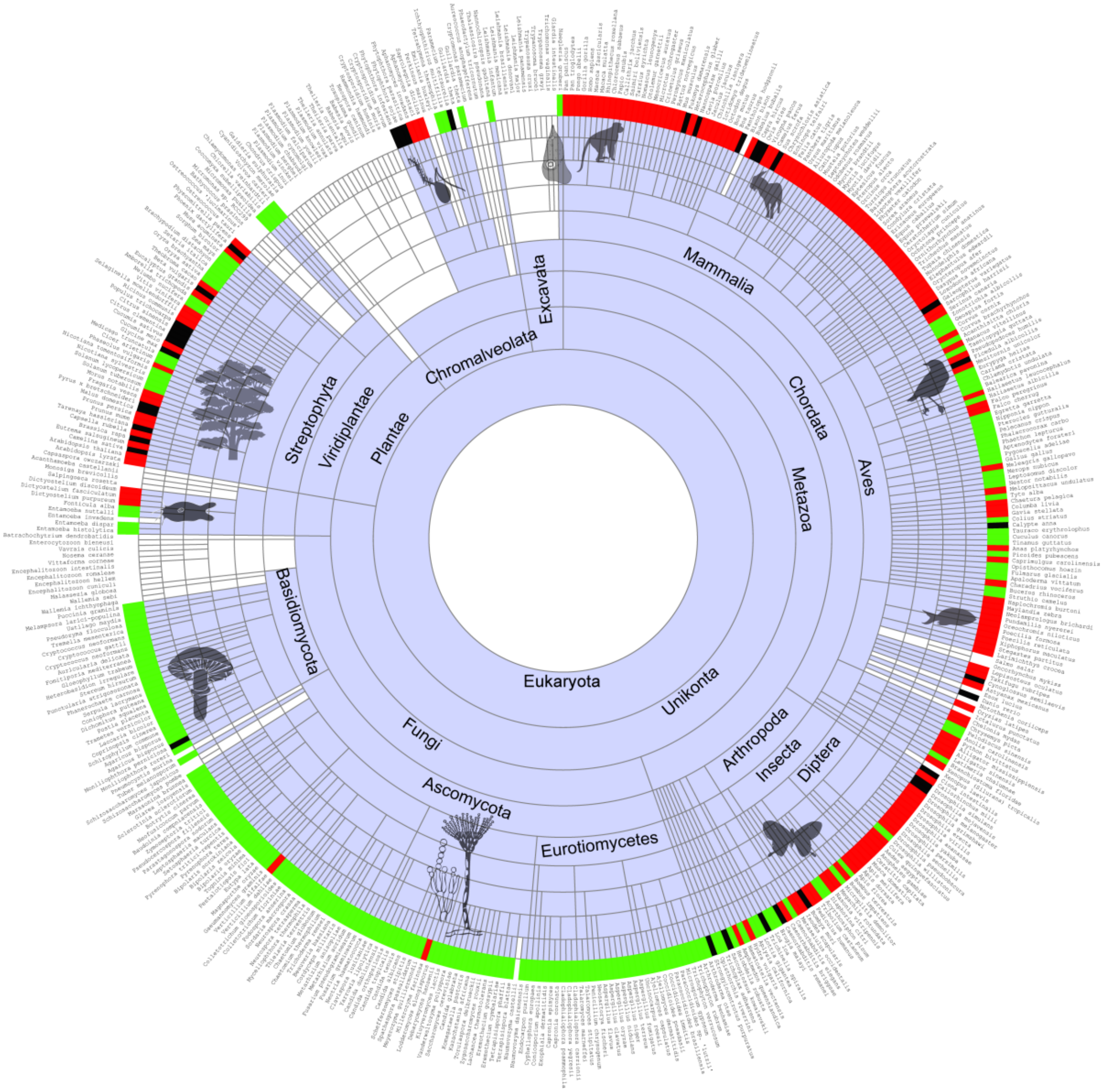


Figure 5-7 Taxonomic distribution of NPC1.

Green flags indicate the presence of a single NPC1 gene in associated genome, whereas red and black flags are representatives of two and more NPC1 copies respectively. Shaded clades have at least one NPC1 gene and white clades have no detectable NPC1 homologs. NPC1 is present in all four eukaryotic supergroups represented here. NPC1 and NPC1-L1 were found in all mammalians except for Bison bison. Although all avian genomes have NPC1, most of them are missing the NPC1-L1 gene. Except for mammals, NPC1-L1 is not consistently found in other classes of Chordata. In Actinopteri, we observed a loss of both genes in Salmoniformes, Esox lucius, Notothenia coriiceps and Ictalurus punctatus. In Anura, Xenopus tropicalis have both genes, but they both are missing from Xenopus laevis.

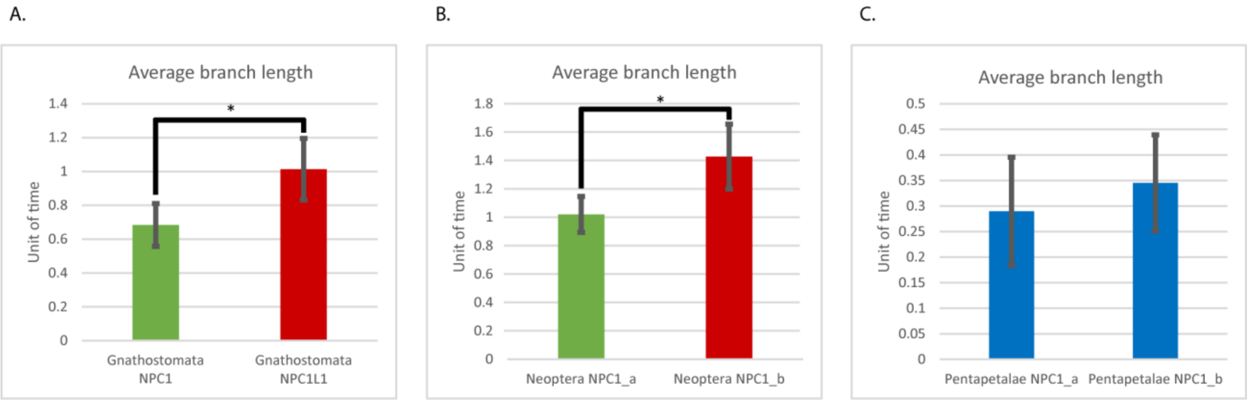


Figure 5-8 Internal diversity of paralogous clades.

Branch length means of clades were measured and subjected to t-test. Significance is represented by asterisk with p-value<0.0001. In C, the p-value was 0.04.

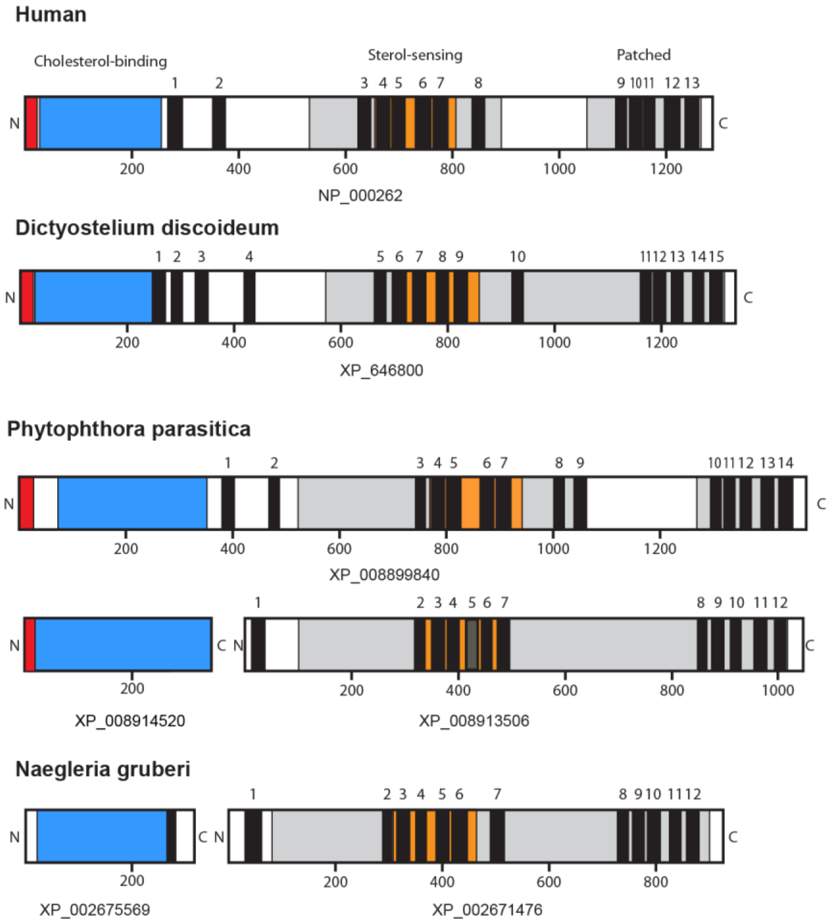


Figure 5-9 Domain architectures of NPC1 homologs.

Full version sequences include cholesterol-binding domain (blue). Short version NPC1 genes contain sterol-sensing and patched domains and no cholesterol-binding domain. Cholesterol-binding domain is found as a single protein in chromalveolatan species as well as Naegleria gruberi. The cholesterol-binding domain protein exclusively coexist with the short version of NPC1. Oomycetes (such as Phytophthora parasitica) have both full and short versions of the NPC1 gene.

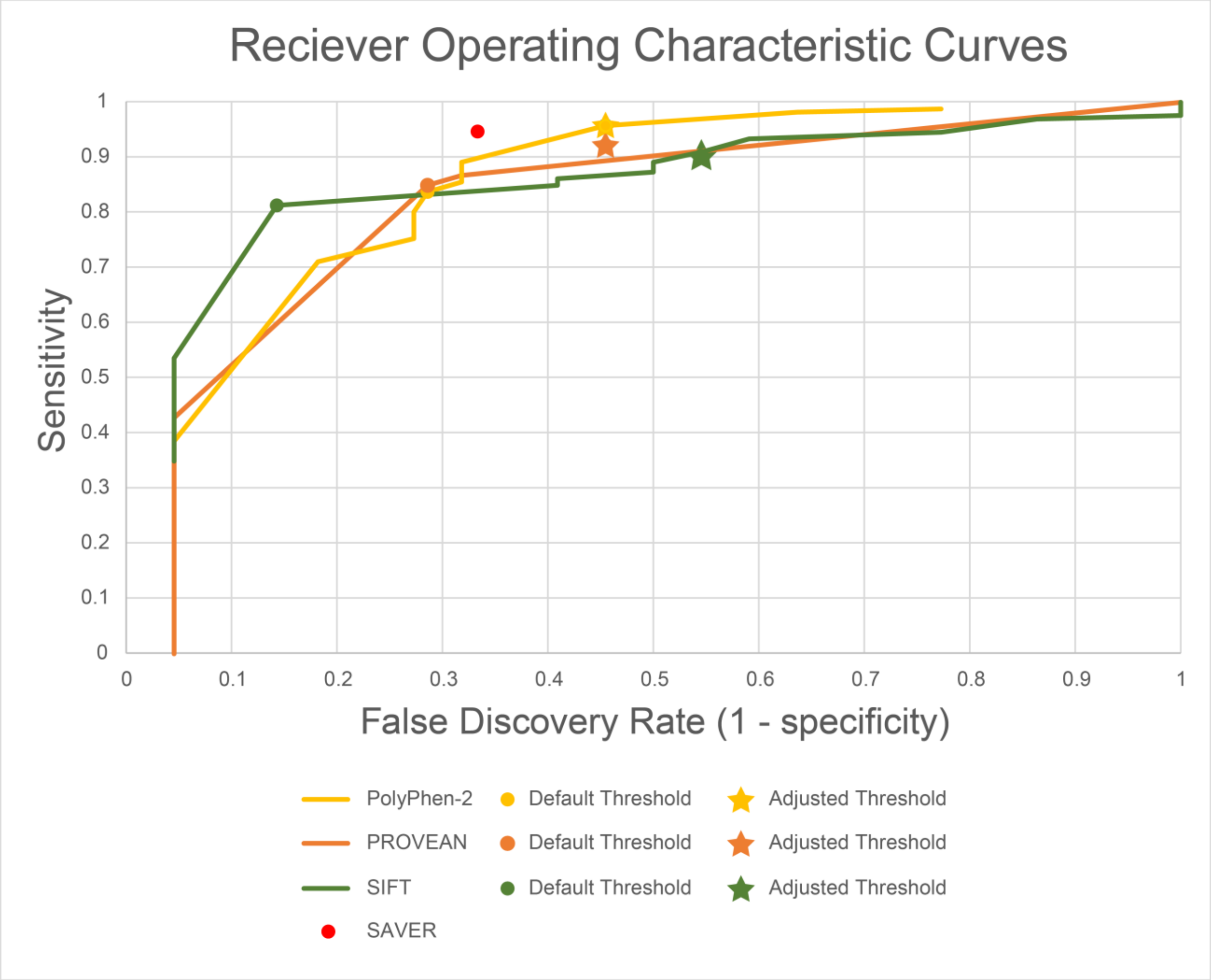


Figure 5-10 Receiver operating characteristics of the tools compared with SAVER datum.

The data is inferred from predictions on 166 damaging and 21 benign variants known in NPC1.

Table 5-2 Known pathogenic and benign mutations in NPC1.

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
A1035V	Pathogenic	D	BB	0.996	D	D	-3.6	D	D	0.002	D	D	(Ribeiro et al. 2001; Fernandez-Valero et al. 2005; Macias-Vidal et al. 2011)
A1054T	Pathogenic	D	BB	1	D	D	-2.85	D	D	0.003	D	D	(Patterson 1993; Millat et al. 2001; Macias-Vidal et al. 2011)
A1062V	Pathogenic	D	BB	0.983	D	D	-3.38	D	D	0.003	D	D	(Millat et al. 2005)
A1132P	Pathogenic	D	BD	0.746	D	D	-2.78	D	D	0.037	D	D	(Mavridou et al. 2014)
A1151T	Pathogenic	D	AA	1	D	D	-3.84	D	D	0	D	D	(Garver et al. 2010; Macias-Vidal et al. 2011)
A1174V	Pathogenic	D	BB	0.978	D	D	-3.22	D	D	0.009	D	D	(Millat et al. 2005)
A1187V	Pathogenic	D	BB	0.999	D	D	-3.2	D	D	0.134	B	D	(Fancello et al. 2009)

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
A1216V	Pathogenic	D	BD	0.997	D	D	-3.22	D	D	0.002	D	D	(Millat et al. 2005)
A388P	Pathogenic	D	BB	0.927	D	D	-3.75	D	D	0.021	D	D	(Park et al. 2003)
A521S	Pathogenic	D	BB	0.003	B	B	-1.12	B	B	0.255	B	B	(Park et al. 2003)
A605V	Pathogenic	D	BD	0.647	D	D	-3.54	D	D	0.005	D	D	(Millat et al. 2001; Sevin et al. 2007)
A745E	Pathogenic	D	BD	0.967	D	D	-4.34	D	D	0.017	D	D	(Park et al. 2003)
A767V	Pathogenic	D	AA	0.997	D	D	-3.82	D	D	0	D	D	(Park et al. 2003)
A926T	Pathogenic	D	BD	1	D	D	-3.82	D	D	0.002	D	D	(Fernandez-Valero et al. 2005; Fancello et al. 2009)
A927V	Pathogenic	D	BB	0.889	D	D	-3.29	D	D	0.006	D	D	(Xiong et al. 2012; Jahnova et al. 2014)
C100S	Pathogenic	D	AA	1	D	D	-9.15	D	D	0	D	D	(Fancello et al. 2009)
C1168Y	Pathogenic	D	AA	1	D	D	-9.7	D	D	0.002	D	D	(Millat et al. 2001)
C177G	Pathogenic	D	AA	0.979	D	D	-11.31	D	D	0	D	D	(Yamamoto et al. 1999)

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
C177Y	Pathogenic	D	AA	0.999	D	D	-10.55	D	D	0	D	D	(Patterson 1993; Ribeiro et al. 2001; Fernandez-Valero et al. 2005; Millat et al. 2005; Macias-Vidal et al. 2011)
C247Y	Pathogenic	D	AA	1	D	D	-10.26	D	D	0	D	D	(Park et al. 2003; Garver et al. 2010)
C479Y	Pathogenic	D	AA	1	D	D	-10.69	D	D	0	D	D	(Fernandez-Valero et al. 2005; Macias-Vidal et al. 2011)
C670W	Pathogenic	D	BD	0.998	D	D	-6.87	D	D	0.001	D	D	(Bauer et al. 2002)
C74Y	Pathogenic	D	AA	1	D	D	-9.93	D	D	0	D	D	(Park et al. 2003; Garver et al. 2010)
C956Y	Pathogenic	D	AA	0.996	D	D	-10.65	D	D	0.001	D	D	(Yamamoto et al. 1999)
D1097N	Pathogenic	D	AA	0.987	D	D	-4.26	D	D	0.062	B	D	(Millat et al. 2005; Sevin et al. 2007)

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
D242H	Pathogenic	D	AA	1	D	D	-6.53	D	D	0	D	D	(Millat et al. 2001)
D700N	Pathogenic	D	AA	1	D	D	-5	D	D	0	D	D	, (Park et al. 2003; Garver et al. 2010)
D874V	Pathogenic	D	BD	0.431	B	D	-5.48	D	D	0.011	D	D	(Millat et al. 2001; Bauer et al. 2002; Kaminski et al. 2002; Vanier and Millat 2003; Sevin et al. 2007; Garver et al. 2010)
D917Y	Pathogenic	D	BB	0.253	B	D	-7.86	D	D	0.001	D	D	(Millat et al. 2005)
D944N	Pathogenic	D	AA	1	D	D	-4.94	D	D	0	D	D	(Millat et al. 2001; Fernandez-Valero et al. 2005; Macias-Vidal et al. 2011; Xiong et al. 2012)
D945N	Pathogenic	D	AA	1	D	D	-4.94	D	D	0	D	D	(Park et al. 2003; Garver et al. 2010)

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
D948H	Pathogenic	D	BB	1	D	D	-6.08	D	D	0.002	D	D	(Fernandez-Valero et al. 2005)
D948N	Pathogenic	B	ZZ	0.998	D	D	-4.14	D	D	0.049	D	D	(Greer et al. 1999; Vanier and Millat 2003; Sevin et al. 2007)
D948Y	Pathogenic	D	BB	1	D	D	-7.99	D	D	0.001	D	D	(Bauer et al. 2002)
E1189G	Pathogenic	D	BB	0.915	D	D	-4.33	D	D	0.008	D	D	(Tarugi et al. 2002)
E451K	Pathogenic	D	BB	0.055	B	D	-1.11	B	B	0.802	B	B	(Tarugi et al. 2002)
E742K	Pathogenic	D	AA	1	D	D	-3.89	D	D	0	D	D	(Park et al. 2003)
F1079S	Pathogenic	D	AA	1	D	D	-7.69	D	D	0	D	D	(Macias-Vidal et al. 2011)
F1087L	Pathogenic	D	AA	0.998	D	D	-5.49	D	D	0.004	D	D	(Park et al. 2003; Garver et al. 2010)
F1224L	Pathogenic	D	AA	1	D	D	-5.38	D	D	0.003	D	D	(Fernandez-Valero et al. 2005; Macias-Vidal et al. 2011)
F537L	Pathogenic	D	BB	0.417	B	D	-4.3	D	D	0.028	D	D	(Millat et al. 2005)

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
F703S	Pathogenic	D	BD	1	D	D	-7.9	D	D	0	D	D	(Yamamoto et al. 2000)
F763L	Pathogenic	D	BD	0.624	D	D	-4.29	D	D	0.007	D	D	(Millat et al. 2005; Sevin et al. 2007)
F779L	Pathogenic	B	ZZ	0.997	D	D	-5.83	D	D	0	D	D	(Bauer et al. 2013)
F995L	Pathogenic	D	AA	0.987	D	D	-5.67	D	D	0.027	D	D	(Macias-Vidal et al. 2011)
G1015V	Pathogenic	D	AA	1	D	D	-8.39	D	D	0.002	D	D	(Yang et al. 2005)
G1034R	Pathogenic	D	AA	0.991	D	D	-5.75	D	D	0.008	D	D	(Yang et al. 2005)
G1140V	Pathogenic	D	AA	1	D	D	-8.03	D	D	0.002	D	D	(Park et al. 2003)
G1209E	Pathogenic	D	AA	1	D	D	-7.05	D	D	0	D	D	(Macias-Vidal et al. 2011)
G1236E	Pathogenic	D	AA	1	D	D	-7.26	D	D	0	D	D	(Yamamoto et al. 2000)
G1240R	Pathogenic	D	AA	1	D	D	-7.26	D	D	0	D	D	(Millat et al. 2005; Runz et al. 2008)
G343E	Pathogenic	D	AA	1	D	D	-7.96	D	D	0	D	D	(Jahnova et al. 2014)
G535V	Pathogenic	D	BD	0.991	D	D	-4.69	D	D	0.009	D	D	(Macias-Vidal et al. 2009)
G640R	Pathogenic	D	BB	1	D	D	-7.77	D	D	0	D	D	(Park et al. 2003)

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
G660S	Pathogenic	D	AA	1	D	D	-6	D	D	0	D	D	(Park et al. 2003; Garver et al. 2010)
G673V	Pathogenic	D	AA	1	D	D	-9	D	D	0.002	D	D	(Park et al. 2003; Garver et al. 2010)
G910S	Pathogenic	B	ZZ	0.995	D	D	-4.8	D	D	0.131	B	D	(Tarugi et al. 2002)
G986S	Pathogenic	D	BB	0.427	B	D	-1.7	B	B	0.342	B	B	(Millat et al. 2001)
G992A	Pathogenic	D	CB	0.017	B	D	-1.85	B	D	0.74	B	B	(Millat et al. 2005; Sevin et al. 2007; Harzer et al. 2014)
G992R	Pathogenic	D	CB	0.789	D	D	-3.06	D	D	0.466	B	B	(Patterson 1993; Millat et al. 2001; Vanier and Millat 2003; Millat et al. 2005; Sevin et al. 2007; Macias-Vidal et al. 2011; Bauer et al. 2013; Jahnova et al. 2014)

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
G992W	Pathogenic	D	BD	0.989	D	D	-3.86	D	D	0.045	D	D	(Patterson 1993; Greer et al. 1998; Greer et al. 1999; Tarugi et al. 2002; Vanier and Millat 2003; Millat et al. 2005; Vanier 2010; Macias-Vidal et al. 2011; Zampieri et al. 2012)
H1016R	Pathogenic	D	AA	0.243	B	D	-6.15	D	D	0.024	D	D	(Park et al. 2003)
H510P	Pathogenic	D	BD	0.988	D	D	-5.2	D	D	0.107	B	D	(Yamamoto et al. 1999)
H512R	Pathogenic	D	AA	0.556	D	D	-6.3	D	D	0.076	B	D	(Bauer et al. 2002; Fancello et al. 2009)

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
I1061T	Pathogenic	D	BD	0.875	D	D	-3.84	D	D	0.001	D	D	(Patterson 1993; Greer et al. 1999; Millat et al. 1999; Millat et al. 2001; Ribeiro et al. 2001; Bauer et al. 2002; Tarugi et al. 2002; Vanier and Millat 2003; Fernandez-Valero et al. 2005; Millat et al. 2005; Sevin et al. 2007; Fancello et al. 2009; Macias-Vidal et al. 2009; Garver et al. 2010; Vanier 2010; Macias-Vidal et al. 2011;

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
													Zampieri et al. 2012; Harzer et al. 2014; Jahnova et al. 2014; Macias-Vidal et al. 2014)
I685T	Pathogenic	D	AA	0.985	D	D	-5	D	D	0.001	D	D	(Zhang et al. 2014)
I943M	Pathogenic	B	ZZ	0.037	B	D	-0.87	B	B	0.226	B	B	(Bauer et al. 2002; Fancello et al. 2009)
K576R	Pathogenic	B	QQ	0	B	B	-1.5	B	B	0.188	B	B	(Fernandez-Valero et al. 2005; Godeiro-Junior et al. 2006)
L1102F	Pathogenic	D	BD	0.999	D	D	-3.4	D	D	0.006	D	D	(Fancello et al. 2009)
L1106P	Pathogenic	D	BD	1	D	D	-5.64	D	D	0.001	D	D	(Macias-Vidal et al. 2011)
L1191F	Pathogenic	D	BD	0.991	D	D	-3.66	D	D	0.003	D	D	(Fancello et al. 2009; Zampieri

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
													et al. 2012)
L1213F	Pathogenic	D	AA	1	D	D	-3.5	D	D	0	D	D	(Yamamoto et al. 1999)
L1213V	Pathogenic	D	AA	1	D	D	-2.62	D	D	0	D	D	(Greer et al. 1999)
L1244P	Pathogenic	D	AA	1	D	D	-5.82	D	D	0.001	D	D	(Runz et al. 2008; Fancello et al. 2009)
L380F	Pathogenic	D	AA	1	D	D	-3.98	D	D	0	D	D	(Park et al. 2003)
L472H	Pathogenic	D	AA	1	D	D	-6.05	D	D	0	D	D	(Fancello et al. 2009)
L648H	Pathogenic	D	BD	0.82	D	D	-5.15	D	D	0.003	D	D	(Fancello et al. 2009)
L684F	Pathogenic	D	AA	1	D	D	-3.83	D	D	0.003	D	D	(Park et al. 2003)
L695V	Pathogenic	D	AA	1	D	D	-3	D	D	0.003	D	D	(Park et al. 2003)
L724P	Pathogenic	D	BD	0.14	B	D	-5.02	D	D	0.001	D	D	(Millat et al. 2001; Sevin et al. 2007)
L929P	Pathogenic	D	BB	0.697	D	D	-2.75	D	D	0.215	B	B	(Park et al. 2003)
M1001T	Pathogenic	D	BB	0.293	B	D	-3.09	D	D	0.005	D	D	(Bauer et al. 2013)
M1142T	Pathogenic	D	BD	0.984	D	D	-5.76	D	D	0	D	D	(Millat et al. 2001; Vanier and Millat

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
													2003; Fancello et al. 2009; Macias-Vidal et al. 2011)
M272R	Pathogenic	D	BD	0.982	D	D	-4.64	D	D	0.001	D	D	(Millat et al. 2001)
M631R	Pathogenic	D	AA	0.999	D	D	-5.87	D	D	0	D	D	(Millat et al. 2001; Millat et al. 2005)
M754K	Pathogenic	D	BD	0.982	D	D	-5.73	D	D	0	D	D	(Fernandez-Valero et al. 2005)
M996R	Pathogenic	D	BB	0	B	B	-2	B	D	0.472	B	B	(Yamamoto et al. 2000)
N1137I	Pathogenic	D	BD	0.99	D	D	-6.93	D	D	0	D	D	(Park et al. 2003)
N1156I	Pathogenic	D	AA	1	D	D	-8.63	D	D	0	D	D	(Fernandez-Valero et al. 2005)
N1156S	Pathogenic	D	AA	1	D	D	-4.8	D	D	0	D	D	(Tarugi et al. 2002)
N188S	Pathogenic	D	BB	0.017	B	D	-1.89	B	D	0.12	B	D	(Bauer et al. 2013)
N222S	Pathogenic	D	BB	0.003	B	B	-1.2	B	B	0.716	B	B	(Park et al. 2003; Fancello et al. 2009)

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
N961S	Pathogenic	B	QQ	0.034	B	D	-2.33	B	D	0.365	B	B	(Jahnova et al. 2014)
N968S	Pathogenic	D	BB	0.034	B	D	-1.8	B	D	0.704	B	B	(Millat et al. 2005; Yang et al. 2005; Fancello et al. 2009)
P1007A	Pathogenic	D	AA	0.995	D	D	-7.78	D	D	0	D	D	(Greer et al. 1999; Millat et al. 2001; Ribeiro et al. 2001; Bauer et al. 2002; Tarugi et al. 2002; Vanier and Millat 2003; Fernandez-Valero et al. 2005; Millat et al. 2005; Godeiro-Junior et al. 2006; Sevin et al. 2007; Fancello et al. 2009; Garver et

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
													al. 2010; Vanier 2010; Macias-Vidal et al. 2011; Zakharova et al. 2012; Zampieri et al. 2012; Bauer et al. 2013; Jahnova et al. 2014)
P166L	Pathogenic	D	AA	0.999	D	D	-8.44	D	D	0.004	D	D	(Millat et al. 2005)
P166S	Pathogenic	D	AA	0.988	D	D	-6.78	D	D	0.006	D	D	(Park et al. 2003; Fancello et al. 2009)
P237L	Pathogenic	D	BD	0.468	D	D	-4.32	D	D	0.041	D	D	(Fancello et al. 2009)
P433L	Pathogenic	D	BD	0.981	D	D	-8.12	D	D	0.001	D	D	(Park et al. 2003; Fancello et al. 2009)
P434L	Pathogenic	B	ZZ	0.001	B	B	0.04	B	B	0.18	B	D	(Fernandez-Valero et al. 2005)

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
P474L	Pathogenic	D	AA	0.889	D	D	-8.08	D	D	0.001	D	D	(Tarugi et al. 2002; Vanier and Millat 2003; Fernandez-Valero et al. 2005; Macias-Vidal et al. 2009; Garver et al. 2010; Xiong et al. 2012; Jahnova et al. 2014)
P543L	Pathogenic	D	AA	1	D	D	-9.6	D	D	0	D	D	(Park et al. 2003; Millat et al. 2005; Garver et al. 2010)
P691L	Pathogenic	D	AA	1	D	D	-10	D	D	0	D	D	(Park et al. 2003; Jahnova et al. 2014)
P733R	Pathogenic	D	BD	1	D	D	-8.75	D	D	0	D	D	(Jahnova et al. 2014)
P867S	Pathogenic	D	BB	0.999	D	D	-7.91	D	D	0.034	D	D	(Park et al. 2003)
P888S	Pathogenic	D	AA	1	D	D	-7.91	D	D	0	D	D	(Fancello et al. 2009)

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
Q775P	Pathogenic	D	AA	1	D	D	-5.4	D	D	0.003	D	D	(Patterson 1993; Millat et al. 2001; Fernandez-Valero et al. 2005; Macias-Vidal et al. 2009; Macias-Vidal et al. 2011)
Q862L	Pathogenic	D	AA	1	D	D	-6.46	D	D	0.002	D	D	(Millat et al. 2005)
Q921P	Pathogenic	D	BD	0.994	D	D	-3.27	D	D	0.098	B	D	(Fancello et al. 2009)
Q92R	Pathogenic	D	BB	0.079	B	D	-2.14	B	D	0.037	D	D	(Ribeiro et al. 2001; Vanier and Millat 2003; Garver et al. 2010)
R1077Q	Pathogenic	D	CB	0.051	B	D	-0.31	B	B	0.063	B	D	(Fancello et al. 2009)
R1186H	Pathogenic	D	BD	1	D	D	-4.74	D	D	0	D	D	(Millat et al. 2001; Vanier and Millat 2003; Millat et al. 2005; Macias-

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
													Vidal et al. 2011; Xiong et al. 2012; Jahnova et al. 2014)
R389C	Pathogenic	D	BD	1	D	D	-7.28	D	D	0	D	D	(Park et al. 2003)
R389L	Pathogenic	D	BB	0.665	D	D	-6.04	D	D	0.058	B	D	(Fancello et al. 2009)
R404P	Pathogenic	D	AA	1	D	D	-6.97	D	D	0	D	D	(Millat et al. 2005)
R404Q	Pathogenic	D	AA	1	D	D	-3.98	D	D	0	D	D	(Millat et al. 2001; Vanier and Millat 2003; Garver et al. 2010)
R404W	Pathogenic	D	AA	1	D	D	-7.96	D	D	0	D	D	(Park et al. 2003)
R411P	Pathogenic	D	BD	0.042	B	D	-1.5	B	B	0.109	B	D	(Jahnova et al. 2014)
R518Q	Pathogenic	B	ZZ	0.02	B	D	-0.27	B	B	0.64	B	B	(Vanier and Millat 2003; Zhang et al. 2014)
R518W	Pathogenic	D	CB	0.997	D	D	-3.7	D	D	0.01	D	D	(Ribeiro et al. 2001)
R615C	Pathogenic	D	AA	1	D	D	-8	D	D	0	D	D	(Park et al. 2003)

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
R615L	Pathogenic	D	AA	0.999	D	D	-7	D	D	0	D	D	(Millat et al. 2005; Sevin et al. 2007)
R726T	Pathogenic	D	AA	0.941	D	D	-5.52	D	D	0.003	D	D	(Zhang et al. 2014)
R789G	Pathogenic	D	BD	1	D	D	-6.81	D	D	0	D	D	(Park et al. 2003; Garver et al. 2010)
R789H	Pathogenic	D	BD	1	D	D	-4.86	D	D	0	D	D	(Zhang et al. 2014)
R934Q	Pathogenic	D	BD	0.199	B	D	-0.79	B	B	0.332	B	B	(Greer et al. 1999; Millat et al. 2001; Vanier and Millat 2003; Millat et al. 2005; Jahnova et al. 2014)
R958L	Pathogenic	D	AA	0.974	D	D	-6.03	D	D	0.005	D	D	(Bauer et al. 2002)
R958Q	Pathogenic	D	AA	1	D	D	-3.25	D	D	0.011	D	D	(Park et al. 2003)
R978C	Pathogenic	D	BD	0.955	D	D	-2.16	B	D	0.05	B	D	(Ribeiro et al. 2001; Vanier and Millat 2003; Di Leo et al. 2004;

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
													Macias-Vidal et al. 2011)
S1200G	Pathogenic	D	AA	0.998	D	D	-3.6	D	D	0.002	D	D	(Bauer et al. 2013)
S1249G	Pathogenic	D	AA	1	D	D	-3.5	D	D	0.008	D	D	(Park et al. 2003; Garver et al. 2010)
S473P	Pathogenic	D	BD	0.709	D	D	-2.46	B	D	0.018	D	D	(Yamamoto et al. 1999)
S636F	Pathogenic	D	BD	0.745	D	D	-4.4	D	D	0.001	D	D	(Fancello et al. 2009)
S666N	Pathogenic	D	BD	0.298	B	D	-2.44	B	D	0.023	D	D	(Jahnova et al. 2014)
S734I	Pathogenic	D	BD	0.998	D	D	-5.78	D	D	0.001	D	D	(Park et al. 2003; Macias-Vidal et al. 2011)
S849I	Pathogenic	D	BD	0.998	D	D	-5.4	D	D	0.001	D	D	(Bauer et al. 2002)
S865L	Pathogenic	D	AA	0.999	D	D	-5.04	D	D	0.002	D	D	(Fernandez-Valero et al. 2005; Millat et al. 2005; Macias-Vidal et al. 2009;

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
													Xiong et al. 2012)
S940L	Pathogenic	D	BD	0.999	D	D	-4.13	D	D	0.005	D	D	(Greer et al. 1999; Vanier and Millat 2003; Millat et al. 2005; Sevin et al. 2007; Garver et al. 2010; Macias-Vidal et al. 2011)
S954L	Pathogenic	D	BD	0.879	D	D	-3.06	D	D	0.123	B	D	(Greer et al. 1999; Bauer et al. 2002; Tarugi et al. 2002; Vanier and Millat 2003; Sevin et al. 2007; Macias-Vidal et al. 2011; Bauer et al. 2013; Jahnova et al. 2014;

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
													Zhang et al. 2014; Maubert et al. 2015)
T1036K	Pathogenic	D	BD	0.997	D	D	-4.09	D	D	0.003	D	D	(Fernandez-Valero et al. 2005)
T1036M	Pathogenic	D	BD	1	D	D	-4.45	D	D	0.001	D	D	(Millat et al. 2005; Garver et al. 2010)
T1066N	Pathogenic	D	BB	0.698	D	D	-3.66	D	D	0.003	D	D	(Fernandez-Valero et al. 2005; Macias-Vidal et al. 2011; Rodriguez-Pascau et al. 2012)
T1205K	Pathogenic	D	AA	1	D	D	-5.3	D	D	0	D	D	(Park et al. 2003; Fancello et al. 2009; Zampieri et al. 2012; Jahnova et al. 2014)

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
T1205R	Pathogenic	D	AA	1	D	D	-5.33	D	D	0	D	D	(Jahnova et al. 2014)
T137M	Pathogenic	D	CB	0.871	D	D	-2.19	B	D	0.04	D	D	(Vanier and Millat 2003; Fernandez-Valero et al. 2005; Garver et al. 2010)
V1023G	Pathogenic	D	AA	0.999	D	D	-6.23	D	D	0	D	D	(Fancello et al. 2009; Zampieri et al. 2012)
V1212L	Pathogenic	D	BB	0.87	D	D	-2.19	B	D	0.011	D	D	(Yang et al. 2005; Xiong et al. 2012; Zhang et al. 2014)
V378A	Pathogenic	D	BB	0.969	D	D	-3.98	D	D	0.001	D	D	(Millat et al. 2001)
V664M	Pathogenic	D	AA	1	D	D	-2.93	D	D	0.001	D	D	(Park et al. 2003; Fernandez-Valero et al. 2005; Macias-Vidal et al. 2011; Bauer et al. 2013; Jahnova

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
													et al. 2014)
V727F	Pathogenic	D	BD	0.87	D	D	-3.51	D	D	0.005	D	D	(Fernandez-Valero et al. 2005)
V780G	Pathogenic	D	BD	0.981	D	D	-6.49	D	D	0	D	D	(Fancello et al. 2009)
V889M	Pathogenic	D	AA	0.995	D	D	-1.82	B	D	0.21	B	B	(Yamamoto et al. 1999; Sevin et al. 2007)
V950G	Pathogenic	D	BD	0.07	B	D	-4.1	D	D	0.003	D	D	(Jahnova et al. 2014)
V950M	Pathogenic	D	BB	0.007	B	B	-1.02	B	B	0.167	B	D	(Millat et al. 2001; Vanier and Millat 2003; Millat et al. 2005; Sevin et al. 2007; Garver et al. 2010)
V959E	Pathogenic	B	QW	0.007	B	B	-2.51	D	D	0.16	B	D	(Fernandez-Valero et al. 2005)
V971L	Pathogenic	D	BD	0.613	D	D	-1.59	B	B	0.234	B	B	(Park et al. 2003)

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
W942C	Pathogenic	D	AA	1	D	D	-12.84	D	D	0	D	D	(Ribeiro et al. 2001; Fernandez-Valero et al. 2005)
Y1019C	Pathogenic	D	AA	1	D	D	-8.37	D	D	0.001	D	D	(Fancello et al. 2009; Zampieri et al. 2012)
Y1088C	Pathogenic	D	AA	1	D	D	-8.64	D	D	0	D	D	(Yamamoto et al. 1999)
Y276H	Pathogenic	D	BD	0.99	D	D	-4.3	D	D	0	D	D	(Jahnova et al. 2014)
Y509S	Pathogenic	D	BD	0.731	D	D	-6.31	D	D	0	D	D	(Park et al. 2003; Garver et al. 2010)
Y634C	Pathogenic	D	AA	1	D	D	-9	D	D	0	D	D	(Fancello et al. 2009)
Y825C	Pathogenic	D	BD	0.997	D	D	-7.75	D	D	0	D	D	(Millat et al. 2001; Vanier and Millat 2003; Millat et al. 2005; Sevin et al. 2007; Garver et al. 2010)

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
Y871C	Pathogenic	D	AA	1	D	D	-8.01	D	D	0.001	D	D	(Millat et al. 2005; Sevin et al. 2007; Macias-Vidal et al. 2011)
Y890C	Pathogenic	D	AA	1	D	D	-8.77	D	D	0	D	D	(Tarugi et al. 2002; Garver et al. 2010)
Y899D	Pathogenic	D	BD	1	D	D	-9.53	D	D	0	D	D	(Tarugi et al. 2002)
A183T	Frequent	B	ZZ	0	B	B	-0.12	B	B	0.586	B	B	(Wassif et al. 2015)
G1073S	Frequent	B	ZZ	0.069	B	D	-1.05	B	B	0.253	B	B	(Wassif et al. 2015)
G911S	Frequent	B	ZZ	0.904	D	D	-4.3	D	D	0.113	B	D	(Wassif et al. 2015)
H215R	Frequent	B	QQ	0	B	B	-1.35	B	B	0.531	B	B	(Wassif et al. 2015)
I450V	Frequent	B	QQ	0.007	B	B	-0.69	B	B	0.182	B	D	(Wassif et al. 2015)
I858V	Frequent	B	QQ	0.047	B	D	-0.72	B	B	0.231	B	B	(Wassif et al. 2015)
M1179V	Frequent	B	QQ	0	B	B	0.32	B	B	0.578	B	B	(Wassif et al. 2015)
M156V	Frequent	D	BB	0.019	B	D	-1.73	B	D	0.067	B	D	(Wassif et al. 2015)
N589S	Frequent	B	QQ	0.001	B	B	-1.01	B	B	0.326	B	B	(Wassif et al. 2015)
P237S	Frequent	B	QQ	0.003	B	B	-2.23	B	D	0.41	B	B	(Wassif et al. 2015)
P434S	Frequent	D	BD	0.001	B	B	-1.34	B	B	0.041	D	D	(Wassif et al. 2015)

Table 5-2 (continued)

Mutation	Pathogenic / Frequent	SAVER	SAVER cat.	PPH2 score	PPH2 (0.452) default	PPH2 (0.012)	PROVEAN score	PROVEAN (-2.5) default	PROVEAN (-1.72)	SIFT score	SIFT (0.05) default	SIFT (0.186)	Reference
Q60H	Frequent	D	CB	0.656	D	D	-2.01	B	D	0.055	B	D	(Wassif et al. 2015)
R1183H	Frequent	D	BD	0.976	D	D	-4.4	D	D	0.002	D	D	(Wassif et al. 2015)
R1266Q	Frequent	B	QQ	0.003	B	B	-0.22	B	B	0.308	B	B	(Wassif et al. 2015)
R411Q	Frequent	B	QQ	0	B	B	-0.15	B	B	0.367	B	B	(Wassif et al. 2015)
R646H	Frequent	B	QQ	0.01	B	B	-1.32	B	B	0.078	B	D	(Wassif et al. 2015)
S1004L	Frequent	B	QQ	0.837	D	D	-3.79	D	D	0.062	B	D	(Wassif et al. 2015)
T511M	Frequent	D	BB	0.978	D	D	-2.88	D	D	0.003	D	D	(Wassif et al. 2015)
V1115F	Frequent	D	CB	0.001	B	B	-2.53	D	D	0.124	B	D	(Wassif et al. 2015)
V810L	Frequent	B	ZZ	0	B	B	-1.09	B	B	0.322	B	B	(Wassif et al. 2015)
W291C	Frequent	D	BD	0.957	D	D	-3.7	D	D	0.074	B	D	(Wassif et al. 2015)

“D” represents damaging and “B” represents benign mutations.

5.9 References for Appendix

Bauer P, Balding DJ, Klunemann HH, Linden DE, Ory DS, Pineda M, Priller J, Sedel F, Muller A, Chadha-Boreham H et al. 2013. Genetic screening for Niemann-Pick disease type C in adults with neurological and psychiatric symptoms: findings from the ZOOM study. *Hum Mol Genet* **22**: 4349-4356.

Bauer P, Knoblich R, Bauer C, Finckh U, Hufen A, Kropp J, Braun S, Kustermann-Kuhn B, Schmidt D, Harzer K et al. 2002. NPC1: Complete genomic sequence, mutation analysis, and characterization of haplotypes. *Hum Mutat* **19**: 30-38.

Di Leo E, Panico F, Tarugi P, Battisti C, Federico A, Calandra S. 2004. A point mutation in the lariat branch point of intron 6 of NPC1 as the cause of abnormal pre-mRNA splicing in Niemann-Pick type C disease. *Hum Mutat* **24**: 440.

Fancello T, Dardis A, Rosano C, Tarugi P, Tappino B, Zampieri S, Pinotti E, Corsolini F, Fecarotta S, D'Amico A et al. 2009. Molecular analysis of NPC1 and NPC2 gene in 34 Niemann-Pick C Italian patients: identification and structural modeling of novel mutations. *Neurogenetics* **10**: 229-239.

Fernandez-Valero EM, Ballart A, Iturriaga C, Lluch M, Macias J, Vanier MT, Pineda M, Coll MJ. 2005. Identification of 25 new mutations in 40 unrelated Spanish Niemann-Pick type C patients: genotype-phenotype correlations. *Clin Genet* **68**: 245-254.

Garver WS, Jelinek D, Meaney FJ, Flynn J, Pettit KM, Shepherd G, Heidenreich RA, Vockley CM, Castro G, Francis GA. 2010. The National Niemann-Pick Type C1 Disease Database: correlation of lipid profiles, mutations, and biochemical phenotypes. *J Lipid Res* **51**: 406-415.

Godeiro-Junior C, Inaoka RJ, Barbosa MR, Silva MR, Aguiar Pde C, Barsottini O. 2006. Mutations in NPC1 in two Brazilian patients with Niemann-Pick disease type C and progressive supranuclear palsy-like presentation. *Mov Disord* **21**: 2270-2272.

Greer WL, Dobson MJ, Girouard GS, Byers DM, Riddell DC, Neumann PE. 1999. Mutations in NPC1 highlight a conserved NPC1-specific cysteine-rich domain. *Am J Hum Genet* **65**: 1252-1260.

Greer WL, Riddell DC, Gillan TL, Girouard GS, Sparrow SM, Byers DM, Dobson MJ, Neumann PE. 1998. The Nova Scotia (type D) form of Niemann-Pick disease is caused by a G3097-->T transversion in NPC1. *Am J Hum Genet* **63**: 52-54.

Harzer K, Beck-Wodl S, Bauer P. 2014. Niemann-pick disease type C: new aspects in a long published family - partial manifestations in heterozygotes. *JIMD Rep* **12**: 25-29.

Jahnova H, Dvorakova L, Vlaskova H, Hulkova H, Poupetova H, Hrebicek M, Jesina P. 2014. Observational, retrospective study of a large cohort of patients with Niemann-Pick disease type C in the Czech Republic: a surprisingly stable diagnostic rate spanning almost 40 years. *Orphanet J Rare Dis* **9**: 140.

Kaminski WE, Klunemann HH, Ibach B, Aslanidis C, Klein HE, Schmitz G. 2002. Identification of novel mutations in the NPC1 gene in German patients with Niemann-Pick C disease. *J Inherit Metab Dis* **25**: 385-389.

Macias-Vidal J, Giros M, Guerrero M, Gascon P, Serratosa J, Bachs O, Coll MJ. 2014. The proteasome inhibitor bortezomib reduced cholesterol accumulation in fibroblasts from Niemann-Pick type C patients carrying missense mutations. *FEBS J* **281**: 4450-4466.

Macias-Vidal J, Gort L, Lluch M, Pineda M, Coll MJ. 2009. Nonsense-mediated mRNA decay process in nine alleles of Niemann-Pick type C patients from Spain. *Mol Genet Metab* **97**: 60-64.

Macias-Vidal J, Rodriguez-Pascau L, Sanchez-Olle G, Lluch M, Vilageliu L, Grinberg D, Coll MJ, Spanish NPCWG. 2011. Molecular analysis of 30 Niemann-Pick type C patients from Spain. *Clin Genet* **80**: 39-49.

Maubert A, Hanon C, Metton JP. 2015. [Niemann-Pick type C disease and psychosis: Two siblings]. *Encephale* **41**: 238-243.

Mavridou I, Cozar M, Douzgou S, Xaidara A, Lianou D, Vanier MT, Dimitriou E, Grinberg D, Vilageliu L, Michelakakis H. 2014. Niemann-Pick type C disease: a novel NPC1 mutation segregating in a Greek island. *Clin Genet* **85**: 543-547.

Millat G, Bailo N, Molinero S, Rodriguez C, Chikh K, Vanier MT. 2005. Niemann-Pick C disease: use of denaturing high performance liquid chromatography for the detection of NPC1 and NPC2 genetic variations and impact on management of patients and families. *Mol Genet Metab* **86**: 220-232.

Millat G, Marcais C, Rafi MA, Yamamoto T, Morris JA, Pentchev PG, Ohno K, Wenger DA, Vanier MT. 1999. Niemann-Pick C1 disease: the I1061T substitution is a frequent mutant allele in patients of Western European descent and correlates with a classic juvenile phenotype. *Am J Hum Genet* **65**: 1321-1329.

Millat G, Marcais C, Tomasetto C, Chikh K, Fensom AH, Harzer K, Wenger DA, Ohno K, Vanier MT. 2001. Niemann-Pick C1 disease: correlations between NPC1 mutations, levels of NPC1 protein, and phenotypes emphasize the functional significance of the putative sterol-sensing domain and of the cysteine-rich luminal loop. *Am J Hum Genet* **68**: 1373-1385.

Park WD, O'Brien JF, Lundquist PA, Kraft DL, Vockley CW, Karnes PS, Patterson MC, Snow K. 2003. Identification of 58 novel mutations in Niemann-Pick disease type C: correlation with biochemical phenotype and importance of PTC1-like domains in NPC1. *Hum Mutat* **22**: 313-325.

Patterson M. 1993. Niemann-Pick Disease Type C. In *GeneReviews(R)*, (ed. RA Pagon, et al.), Seattle (WA).

Ribeiro I, Marcao A, Amaral O, Sa Miranda MC, Vanier MT, Millat G. 2001. Niemann-Pick type C disease: NPC1 mutations associated with severe and mild cellular cholesterol trafficking alterations. *Hum Genet* **109**: 24-32.

Rodriguez-Pascau L, Toma C, Macias-Vidal J, Cozar M, Cormand B, Lykopoulou L, Coll MJ, Grinberg D, Vilageliu L. 2012. Characterisation of two deletions involving NPC1 and flanking genes in Niemann-Pick type C disease patients. *Mol Genet Metab* **107**: 716-720.

Runz H, Dolle D, Schlitter AM, Zschocke J. 2008. NPC-db, a Niemann-Pick type C disease gene variation database. *Hum Mutat* **29**: 345-350.

Sevin M, Lesca G, Baumann N, Millat G, Lyon-Caen O, Vanier MT, Sedel F. 2007. The adult form of Niemann-Pick disease type C. *Brain* **130**: 120-133.

Tarugi P, Ballarini G, Bembi B, Battisti C, Palmeri S, Panzani F, Di Leo E, Martini C, Federico A, Calandra S. 2002. Niemann-Pick type C disease: mutations of NPC1 gene and evidence of abnormal expression of some mutant alleles in fibroblasts. *J Lipid Res* **43**: 1908-1919.

Vanier MT. 2010. Niemann-Pick disease type C. *Orphanet J Rare Dis* **5**: 16.

Vanier MT, Millat G. 2003. Niemann-Pick disease type C. *Clin Genet* **64**: 269-281.

Wassif CA, Cross JL, Iben J, Sanchez-Pulido L, Cougnoux A, Platt FM, Ory DS, Ponting CP, Bailey-Wilson JE, Biesecker LG et al. 2015. High incidence of unrecognized visceral/neurological late-onset Niemann-Pick disease, type C1, predicted by analysis of massively parallel sequencing data sets. *Genet Med* doi:10.1038/gim.2015.25.

Xiong H, Higaki K, Wei CJ, Bao XH, Zhang YH, Fu N, Qin J, Adachi K, Kumura Y, Ninomiya H et al. 2012. Genotype/phenotype of 6 Chinese cases with Niemann-Pick disease type C. *Gene* **498**: 332-335.

Yamamoto T, Nanba E, Ninomiya H, Higaki K, Taniguchi M, Zhang H, Akaboshi S, Watanabe Y, Takeshima T, Inui K et al. 1999. NPC1 gene mutations in Japanese patients with Niemann-Pick disease type C. *Hum Genet* **105**: 10-16.

Yamamoto T, Ninomiya H, Matsumoto M, Ohta Y, Nanba E, Tsutsumi Y, Yamakawa K, Millat G, Vanier MT, Pentchev PG et al. 2000. Genotype-phenotype relationship of Niemann-Pick disease type C: a possible correlation between clinical onset and levels of NPC1 protein in isolated skin fibroblasts. *J Med Genet* **37**: 707-712.

Yang CC, Su YN, Chiou PC, Fietz MJ, Yu CL, Hwu WL, Lee MJ. 2005. Six novel NPC1 mutations in Chinese patients with Niemann-Pick disease type C. *J Neurol Neurosurg Psychiatry* **76**: 592-595.

Zakharova E, Mikhailova SV, Proshliakova T, Rudenskaia GE. 2012. [Clinical and genetic special features of Niemann-Pick disease, type C]. *Vestn Ross Akad Med Nauk*: 60-65.

Zampieri S, Bembi B, Rosso N, Filocamo M, Dardis A. 2012. Treatment of Human Fibroblasts Carrying NPC1 Missense Mutations with MG132 Leads to an Improvement of Intracellular Cholesterol Trafficking. *JIMD Rep* **2**: 59-69.

Zhang H, Wang Y, Lin N, Yang R, Qiu W, Han L, Ye J, Gu X. 2014. Diagnosis of Niemann-Pick disease type C with 7-ketocholesterol screening followed by NPC1/NPC2 gene mutation confirmation in Chinese patients. *Orphanet J Rare Dis* **9**: 82.

CHAPTER 6. Conclusion

The main subjects of this work are protein sequences. Three aspects of proteins are covered here: (i) Identifying protein domain architectures; (ii) Finding functional partners of proteins; (iii) Evaluating importance of individual residues in protein function. With the increasing amount of genomic data, it is becoming more important to computationally identify cellular and molecular functions of proteins by using only sequences. To define a protein, looking into the domains that the protein contains is one of the initials steps. More often than not, researchers encounter proteins with no assigned domain for some regions, if not for the entire length of the protein. That's why more sensitive domain assignment was a necessary task (chapter three). Secondly, while using the comparative genomics approach, obtaining the taxonomic distribution of the gene/protein of interest is another useful aim. However, displaying the distribution of sequences only in species where a particular protein is present constitutes only half of the story. Information on both presence and absence of proteins completes the phylogenetic profile. Additionally, visualizing multiple proteins on the same distribution frame is helpful to understand their coevolution patterns, if there are any. By introducing such features in the field of bioinformatics, it becomes possible to learn more about evolutionary history and functional interactions of proteins (chapter four). Finally, identifying amino acids that are functionally important in a protein was an important task as it has a wide applicability in research and clinics. A general approach is to examine the conservation patterns of individual amino acids. However, with this traditional approach does not provide the desired level of accuracy (mostly because of insensitivity) in predicting damaging variants. This is likely due to the noise introduced by the non-related or neo-functionalized paralogous proteins, which were likely to have the same function as the protein of interest.

Analyzing evolutionary history thoroughly with the phylogenetic approach helped in distinguishing between diverged paralogs and orthologs. By considering only orthologs, the quality of the alignment increases, yielding a better assessment of conserved residues in the MSA of functionally equivalent sequences (chapter five).

6.1 Implications of the Covered Work for the Present and the Future

6.1.1 Protein domain assignment

The first work in this dissertation was building the CDvist algorithm and web server. The goal of developing CDvist was to accelerate the discovery of novel domains and improving overall domain coverage in protein sequences. Despite the wide popularity and success of domain identification tools and databases, as of today, at least 30% of the protein sequence space (all sequences in the non-redundant database) has no domain coverage. By providing a logical, flexible and iterative pipeline for domain search, rich visualization, and bulk querying, we expected CDvist to attract both biologists and bioinformaticians. As of January 2015, since the manuscript was published, more than 3000 users visited the CDvist website with ~4000 sessions. CDvist has become of interest to researchers dealing with proteins with no identified domain with standard techniques. Also, users having batch queries use the tool to perform more sensitive domain identification through HHsearch on multiple queries at once. With the help from users contacting us, we tailored the web server according to their needs. We expect that this kind of integration of ideas helps a variety of research groups. For this reason, it is important that administrators of this kind of research web server should be accessible, responsive and prepared to fulfill the needs of the community.

CDvist works on subsequences rather than the entire sequence. Subsequences are redefined after every iteration that comes with a significant hit. Performing HHsearch with a number of small units takes time and that’s why the CDvist process is slow for a single sequence submission. The tool compensates this disadvantage by offering linear speed-up for batch requests using a supercomputer, Newton High-Performance Cluster in the University of Tennessee. In other words, CDvist users all around the world utilize a supercomputer from their home.

Users of CDvist are offered a number of options for domain search. It is possible to manually choose the databases of interest in the desired order. Also, since the best domain hit (if it is beyond the significance threshold) is assigned to the protein each time, the domain assignment and the following subsequence definition depend heavily on the determined significance threshold. The order of the databases changes the entire result of domain architecture. If a user starts with Pfam followed by CDD, it means that the user gives more importance to Pfam and refers to CDD only to find domains for the regions that couldn’t be covered by Pfam. The domain architecture found first by CDD followed by Pfam would be different in a sense that the order of tools would be opposite to the case above. For this reason, there is no standard procedure to search protein domains with CDvist. The pipeline can be tailored according to each user’s need. That’s why pre-computed CDvist results cannot be offered for now. If a concept that rationalizes a standard procedure is developed, then that CDvist pipeline can be applied to a vast amount of protein data. Only if there is such a standardized procedure, it would be reasonable to dedicate computer hours to perform CDvist search on the protein universe, after which pre-computed domain assignments would be retrieved at a high speed.

It is predicted that, if the domain discovery rate stays still, domain databases will cover the complete protein universe in the next 20-30 years. Until that point, CDvist is likely to help in this effort.

6.1.2 Using the Positives and Negatives of Multidimensional Data

Humans have extraordinary ability in pattern recognition. We use this ability to conceive the facts. In molecular biology, although the data is concrete, it is not always easy to visualize. Especially for large sets of genomic data, researchers tend to cluster data and display representatives instead of visualizing everything. With Aquerium, we offer a platform for visualization of multi-layer genomic information. Among comparative genomics methods, particularly phylogenetic profiling is appropriate to be applied on such a platform displaying thousands of genomes with their multi-layer information.

One important missing piece in visualizing molecular data on genomics level is to display negative data. Since it is challenging to display what is absent in genomes, it has been ignored by the bioinformatics community. Aquerium handles genomic absence information. Although single absences are not confidently categorized as true, multiple independent co-absences are important indicators of true negatives.

Aquerium uses the taxonomic information to cluster related organisms. Although this classification is useful, it is limited in identifying relationships between taxons which have no hierarchical discrimination between them. For instance, relationships between species belonging to the same genus cannot be determined if there is no middle rank between the species and the genus rank. Therefore, presence and absence data in these kinds of genomes may misleadingly seem as independent events. To solve this problem, a phylogenetic tree of life should be built to guide neighboring genomes that are closer to each other. However, building the tree of life is not straightforward, and there is a substantial effort on this subject in computational biology.

Phylogenetic profiling is usually applied for genes/proteins that are suspected to be involved in the same biochemical network. This method is used to verify protein-protein interactions as an additional line of evidence from the genomics point of view. However, this approach can also be used as a hypothesis generating method in discovering potentially interacting genes/proteins by building phylogenetic profiles of proteins and domains and compare them with each other. Aquerium already contains a database of domain-linked genomes. This database can easily be converted to a genome-linked domain database. This database should include domain records, each of which contains a genomic occurrence pattern (profile). Because taxonomic classification can provide a sensitivity up to a certain limit, the phylogenetic tree of life will be crucial when building precise profiles. Phylogenetic profiles of domains can be compared to each other to reveal common co-occurrence patterns which would be the indicators of potentially interacting domains. Using only Pfam (v.28) domains, the analysis would result in 256 million pairwise comparisons, which is a solvable problem in computational biology.

6.1.3 Phylogenetics Matter in Health and Disease

Predicting the effects of mutations in genetic diseases is one of the hot topics of computational biology, with direct applications in personalized medicine, risk estimation etc. We brought up a new concept in this prediction algorithm: using precise evolutionary histories of genes. Building high-quality evolutionary history is challenging as there is no standard automated procedure to establish it. However, every automation begins with primary concepts manually proved to work in test cases. In chapter five, we present such a test case with new ideas on how to assess single amino acid variants. The major point of the work was to distinguish orthologs from paralogs. We defined criteria on how to perform such a task through phylogenetic analysis. Also, new functions in isoform handling and MSA cleaning were introduced. Moreover, a new evolutionary parameter "event number" was defined and used in the algorithm. We think that these new approaches should be taken into consideration by the community dealing with mutation outcome prediction.

Although the introduced concepts worked for the *NPC1* case, it should be tested in other Mendelian diseases as well. Especially, the genes with different patterns of evolutionary history should be tested. As a first step, our manual work can be performed on ~600 well-known Mendelian diseases. Phylogenetic trees for this set should be built and analyzed. After testing the genes with diverse evolutionary patterns, it would be possible to automate the approach and algorithm to apply to any protein. The SAVER algorithm should be exposed to machine learning and trained with a benchmark data set. By training the algorithm, SAVER can produce a continuous spectrum of scores between benign and damaging instead of binary outputs. This way, the results will be biologically more meaningful because some mutations may have subtle effects where others may result in function loss or neutral change.

As a side product, the phylogenetic analysis was proved useful in the assignment of orthologs and paralogs. For this reason, phylogenetics should be adopted by orthology databases. The current methods in these databases are generally based on pairwise comparisons and/or MSAs converted into distances. However, as discussed in the introduction, phylogenetic analysis provides a better resolution in discriminating orthologs from paralogs.

6.2 Scripting for Genomics

This work resulted in a number of scripts, packages and databases as side products. A small portion of them was polished and made available to public with a user-friendly graphical interface. However, most of them still remain in private depositories and are not available to the public. Unpublished scripts produced in this work were not crucial for final products of the projects, however, they were very useful in working with genomic data. In an ideal world, researchers should benefit from each other’s work, and programming code to make progress in science. Reinventing the wheel would be a waste of time, money and labor. Therefore, as computational biologists, we should make our resources available to anyone. GitHub and BitBucket are appropriate repository hosting services.

JavaScript/HTML5 is a good scripting platform for not only presenting data to public but also understanding genomic data for in-house usage. Jquery and D3 are game-changer JavaScript packages as they established new ways of coding. Moreover, browser extensions written in JavaScript are highly useful to add external lines of information on the existing web servers such as BLAST. These packages should be made available to the public. However, the small visualization packages are often not worth to prepare for publishing. I believe that there should be a journal dedicated to small application notes for genomic data visualization. Therefore, researchers who build their in-house applications are encouraged to share them with the community.

6.3 Final Remarks

This dissertation showed how considering molecular evolution helps in understanding the functions of the proteins overall and on the residue-level. Diversity is achieved through evolution while conserving the most important parts for survival at the molecular level. Genes are derived from each other and all genes could have had a single common ancestor. Although sequence data are too diverged to achieve the universal common ancestor, it is still useful to infer relatively recent relationships among them. This homology information helps greatly in understanding molecular biology, eventually leading to a solid understanding of cellular and even organismal levels of life. Through evolution, genes that are not viable have already been eliminated. Therefore, what we observe in the evolution of genes and organisms can be used as a guide to understanding what changes were allowable.

There is still a lot to be discovered from sequences that are already available. Because sequence data contains more than one type of information, a variety of techniques are applied to understand more about gene functions, protein structures, disease tendency etc. Every method/approach developed in this field has different aims and priorities. In other words, there is no such an optimal algorithm for a biological question. Every different case is evaluated afresh by computational biologists. While biologists can use bioinformatics tools for simple tasks, a stronger computational expertise is required to analyze large data sets. To conclude, in order to perform large-scale bioinformatics analysis, computational programs, resources and power are necessary but insufficient unless they are performed by a computational biologist.

VITA

Ogun Adebali was born in Izmir, Turkey. He attended Izmir Ataturk High School. He received his Bachelor of Science degree in Molecular Biology and Genetics from Middle East Technical University, Ankara, Turkey in 2011. In the last year of the university, he started to get interested in computational work. Then, he got married and moved to Knoxville, Tennessee. He was enrolled in the Graduate School of Genome Science and Technology which is a joint program between the University of Tennessee and Oak Ridge National Laboratory. After joining Igor Jouline’s Computational Genomics research group, he became a "dry-lab" researcher, which he enjoyed a lot. He likes acting, dancing, choreographing and playing poker.