Evaluation of in silico databases for the classification of genomic alterations in oncological samples

Amann, Veronique Agnes

UNIVERSITY OF SPLIT
SCHOOL OF MEDICINE

Véronique Amann

EVALUATION OF *IN SILICO* DATABASES FOR THE CLASSIFICATION OF GENOMIC ALTERATIONS IN ONCOLOGICAL SAMPLES

Diploma thesis

Academic year:
2021/2022

Mentor:
Prof. Johannes Brachmann, MD, PhD

Coburg, August 2022

UNIVERSITY OF SPLIT
SCHOOL OF MEDICINE

Véronique Amann

EVALUATION OF *IN SILICO* DATABASES FOR THE CLASSIFICATION OF GENOMIC ALTERATIONS IN ONCOLOGICAL SAMPLES

Diploma thesis

Academic year:
2021/2022

Mentor:
Prof. Johannes Brachmann, MD, PhD

Coburg, August 2022

TABLE OF CONTENTS

I would sincerely like to thank my mentor Prof. Brachmann for making this final work possible.

Furthermore, I would like to thank the team of the department of pathology. Especially Prof. Aigner, for offering prompt help, whenever I needed it.

I would also like to thank Ms. Gaudiello warmly, for offering her great organizational talent and help throughout this project.

Finally, I would like to express my heartfelt thanks to my family and close friends, for always supporting me.

LIST OF ABBREVIATIONS

DNA - deoxyribonucleic acid

RNA - ribonucleic acid

HGP - human genome project

BRCA1 - breast cancer gene 1

BRCA2 - breast cancer gene 2

EGFR - epidermal growth factor receptor

KRAS - kirsten rat sarcoma viral oncogene homolog

HPV - human papillomavirus

NGS - next-generation sequencing

GRCh38 - genome research consortium human build 38

HGVS - human genome variation society

VCF - variant call format

ACMG - american college of medical genetics and genomics

VUS - variant of unknown significance

SNP - single-nucleotide polymorphism

nsSNP - non-synonymous single nucleotide polymorphisms

MCC - matthews correlation coefficient

PMID - pubmed identifier

1. INTRODUCTION

1.1 Genome

The entirety of the total genetic information of an organism allowing it to function - this is the genome. In living beings, this information is retained in the DNA (deoxyribonucleic acid) in form of chromosomes. Genes are small parts of DNA coding for RNA (ribonucleic acid) and proteins, that are needed by the organism for its existence. The transcriptome is the expression of all RNA molecules encoded by the genome and the proteome is the full expression of its proteins. There is a difference between eukaryotes and prokaryotes regarding the location of storage of the genome. Former carry their genome in the nucleus, which is a cell component surrounded by a membrane, whereas in the latter, the genome floats freely, without a membrane, in the cytoplasm - the nucleoid. The science of exploring genomes is called genomics. (1)

The Human Genome Project (HGP) was a universal collaboration of scientists, aiming to decipher the complete genome of the species Homo Sapiens. The project started in 1990 and finished 13 years later, in 2003. In the beginning, chromosomes were split into large overlapping segments. Sequencing and alignment of these segments were then performed. The parts, that were still left were also sequenced at the end. Decoding every human gene in the DNA and making the results available for everyone was a big milestone for further research in many scientific areas (2).

In molecular medicine, for example, sequencing the whole human genome made diagnosing and treating gene-linked diseases on a DNA level possible. Understanding diseases at the site of origin serves as a great base for improved personalized medicine and disease prevention (3).

Even though, the HGP was seen as an epiphany of the complete human genome, around 15 % were not sequenced because of restrictions in technology. During the next decade, researchers could reduce this number to 8%. By applying state-of-the-art technology in sequencing, it was possible to complete the missing sequences of the genome. Scientists used a complete hydatidiform mole, an anucleated ovum fertilized by a sperm, for this task. After the fusion of the germ cells, the evolving cell comprised only a paternal chromosome, which made sequencing for scientists easier, since a differentiation between the chromosomes was not necessary anymore. In the end, 3,05 billion DNA base pairs were sequenced in total, with 0,3% of the genome still carrying some inaccuracies. Rectifying these sequences enabled researchers for having a new mission in the future (4) (Figure 1).

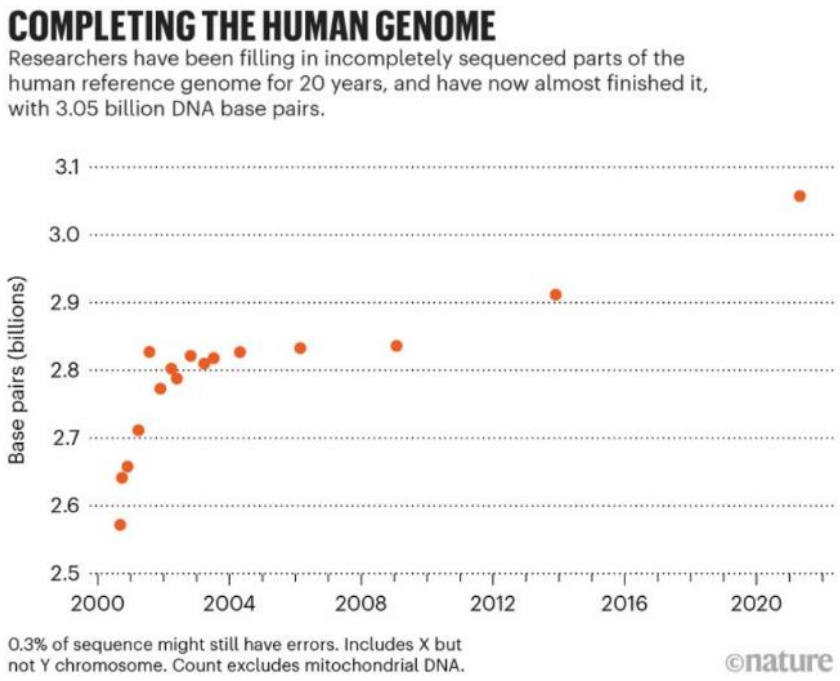

Figure 1. Timeline of the Human Genome Project

Source: https://www.nature.com/articles/d41586-021-01506-w

1.2 Proteins

"Proteins serve a variety of functions within cells. Some are involved in structural support and movement, others in enzymatic activity, and still others in interaction with the outside world. Indeed, the functions of individual proteins are as varied as their unique amino acid sequences and complex three-dimensional physical structures (5)."

Amino acids are the basic component of proteins. Carbon is in the center of each amino acid, attached to it, are a hydrogen-, carboxyl-, and an amino group, plus an R group, that can vary. The R group determines to which class of amino acids it belongs (Figure 2) (6).

The individual formation of every protein is defined by the genetic code of each gene. This leads to a unique sequence of amino acids and distinct bonds, that are responsible for many different "three-dimensional structures" or "conformations" (7).

The primary structure of a protein describes the sequence of amino acids in a chain. This sequence is fixed and corresponds to the information of the gene, that codes for the corresponding protein (8).

The amino acid chain has, because of peptide bonds, many freely lying keto- and amino groups. These groups form interaction with each other, leading to the formation of the

secondary structure. The sidechains of the amino acids are not involved in this process. The most common secondary structures are called α-helix and β-sheet (8).

The tertiary structure describes the three-dimensional structure of a protein and develops through torsion of the secondary structure. Now, the side chains of the amino acids form interactions and these are stabilized by covalent and non-covalent connections (disulfide bridges, hydrogen bonds, ionic interactions, hydrophobe interactions) (8).

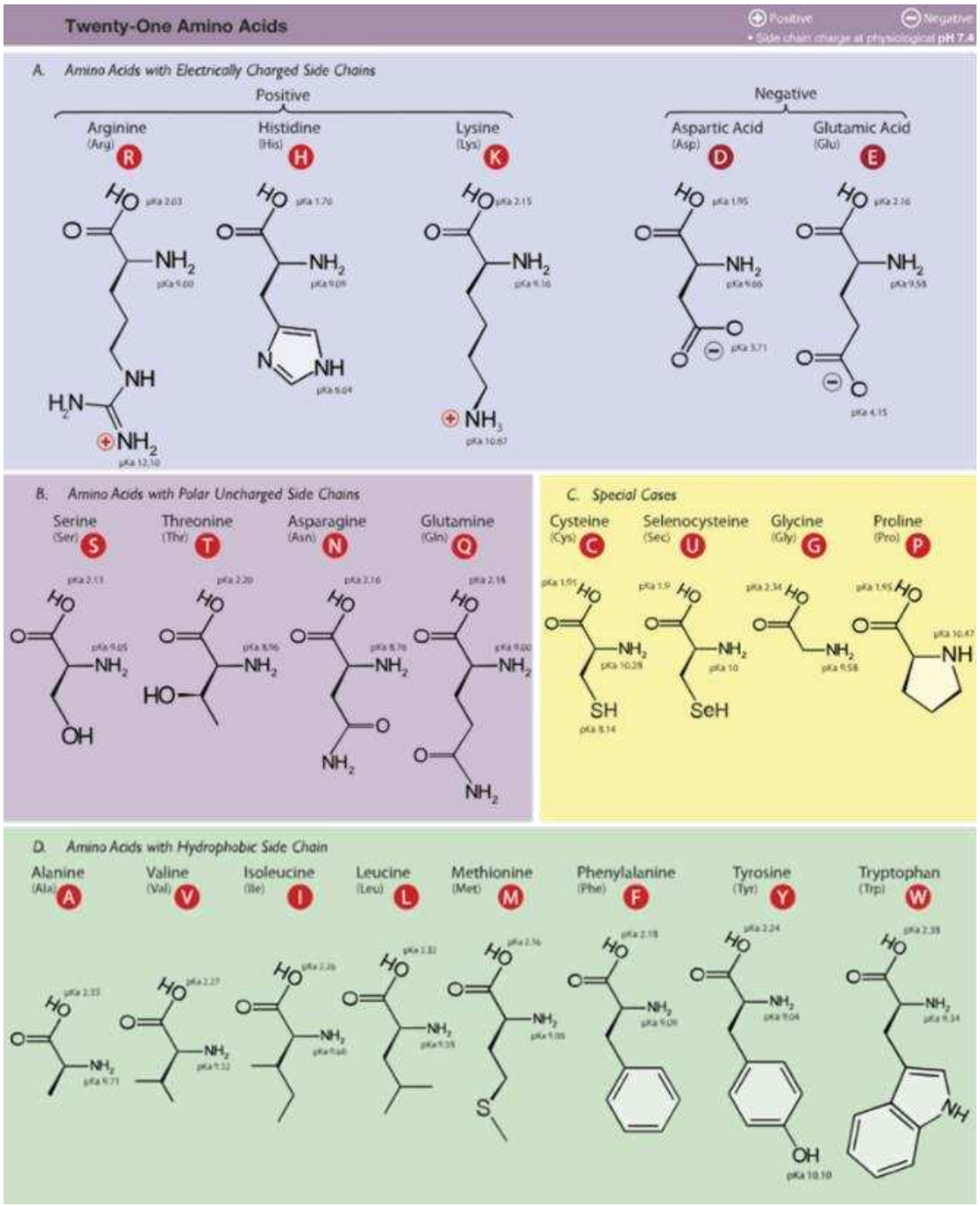

Figure 2. Subgroups of proteinogenic amino acids

Source: https://upload.wikimedia.org/wikipedia/commons/a/ac/AAs_table.png

The quaternary structure describes symbioses of proteins. Many three-dimensional units form together a bigger functional unit. Enzyme complexes, ribosomes, or protein fibers are examples of such supra-molecular structures (8).

1.3 Mutations

A mutation is “any change in the DNA sequence of a cell” (9). Errors during cellular division serve as intrinsic factors to the emergence of mutations whereas ionizing radiation, chemicals or viruses can lead as DNA-harming, extrinsic factors to the development of mutations. Mutations can be categorized as either damaging, favorable or neutral. Germline mutations take place in egg- and sperm cells and thus can be transmitted to the next generation, whereas in somatic mutations no risk for heritage is present. Diseases, such as cancer, can follow different mutations. A mutation with an unknown impact can also be called a variant (8).

1.3.1 Types of mutations

Mutations are categorized into three main types:

- genome mutation
- chromosome mutation
- gene mutation (10)

Each animal has its characteristic number of chromosomes. In humans, this would be a set of 23 chromosomes twice. In a genome mutation, also called numerical chromosomal aberration, the number of chromosomes in a cell is changed. If the complete set of chromosomes is multiplied, it is called polyploidy. If single chromosomes are multiplied or missing, it is called aneuploidy. Most genome mutations cannot be survived by humans. The main cause of genome mutations is a non-disjunction during mitosis or meiosis. Down syndrome, Turner syndrome and Klinefelter syndrome are known diseases caused by aneuploidy in the germline (11).

A chromosome mutation, also called structural chromosomal aberration, is a change in the structure of one or more chromosomes. The genetic information is disrupted or rearranged. Depending on how many genes are affected by this change, it can lead to more or less serious complications.

Chromosome mutations can be divided into:

- Deletion
- Duplication
- Insertion
- Inversion
- Translocation
- Formation of isochromosomes

Chromosome mutations are often inherited from one parent, that already has an existing mutation or they can occur because of mistakes during meiosis, especially during the crossing over. Well known examples are the cri-du-chat syndrome or in leukemia patients the Philadelphia chromosome (12).

A gene mutation is the change of the genetic information in a gene. It could lead to an altered sequence of nucleotides, which can have an impact on protein synthesis. Gene mutations are point mutations or frameshift mutations, that potentially lead to a stop codon (13).

A point mutation is an alteration in a single base. Therefore, this could lead to one different amino acid during translation. It is caused by a substitution. One base is substituted by another base. The substitution can be further subdivided into a transition or a transversion (13).

In a transition, one purine base is exchanged for another purine base (Adenine, Guanine). Whereas in a transversion a purine base gets exchanged with a pyrimidine base (Cytosine, Thymine, Uracil) or vice versa. Depending on which codon results from the 3 successive bases, the point mutation can be further divided into a silent mutation, a missense mutation or a nonsense mutation (13).

In a silent mutation, the codon still codes for the same amino acid, even though one base has changed. In a missense mutation, the codon codes for a different amino acid and in a nonsense mutation the codon does not code for an amino acid and leads to a stop codon (13).

In a frameshift mutation, one or more bases can be inserted (insertion) into the sequence of the DNA or deleted (deletion), leading to a shift in the sequence that is following, resulting in different amino acids and different proteins. If only one base is inserted or deleted, it is also often assigned to a point mutation. In the case of an insertion or deletion of three bases or a multiple of three, only one or a few proteins are affected (13).

For example, a mutation in the BRCA1 or BRCA2 gene, that increases the risk of developing breast or ovarian cancer (14).

1.3.2 Effects of mutations

Another way of categorizing mutations is through the consequences for an organism. It can influence the structure and function of proteins. This can also influence the phenotype of an organism (10). Effects of mutations can be:

- Silent mutation
- Neutral mutation
- Loss-of-function mutation
- Gain-of-function mutation
- Conditional mutation
- Lethal mutation
- Mutations as an evolutionary factor

A silent mutation does not have an influence on the structure or function of a protein. This is possible for example, if the mutation is in the non-coding part of the DNA or if a different base sequence still leads to the same amino acid, because of the redundancy of the genetic code. This is also called a synonymous mutation. There is mostly more than one codon that codes for an amino acid. This can be well seen in the genetic code chart (10) (Figure 3).

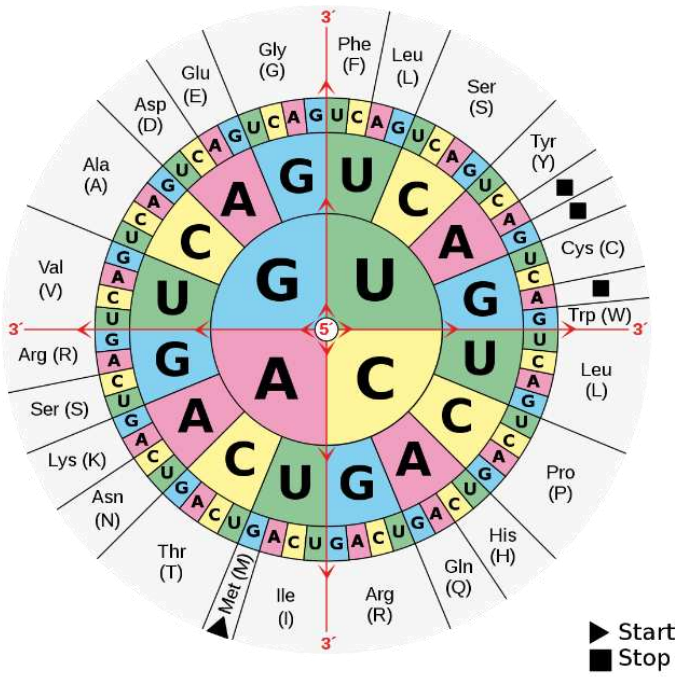

Figure 3. Genetic code chart

Source: https://de.wikipedia.org/wiki/Genetischer_Code

A neutral mutation leads to a different amino acid sequence in a protein, but it has no effect on the function of the protein. In variable regions of a protein, an exchange of proteins or a gab is possible, without leading to a loss of function. The protein, that has a substituted amino acid, still belongs to the same protein group (10).

A loss-of-function mutation is often a frameshift or missense mutation. It leads to a protein, which has lost its function. A mutation in the BRCA1 or BRCA2 gene is an example of this type of mutation (10).

In a gain-of-function mutation, the produced protein has an expanded function. This is the case in some types of cancer when the mutated protein leads to excessive cell growth. For example, a mutation in the EGFR in lung cancer or the mutation in the KRAS gene in colon cancer (10).

A conditional mutation can be a cause of different phenotypes. One common condition is temperature. In the organism, the enzyme protein is only changed under a certain temperature. This can be well-observed in some kinds of rabbit breeds. A mutated gene is responsible for black fur, which is inactive during high temperatures. Rabbits have light fur, because of the body temperature. Only the extremities are covered with black fur (10).

A lethal mutation leads always to the death of the organism. For example, mutations in the human germline (10).

Mutations do not only have neutral or negative consequences. They also contribute to biodiversity on earth, through a change in the genome, which causes positive results in the organism. If this mutation carries on through natural selection, it leads to greater biological diversity. Humans for example, that are affected by sickle cell anemia, are protected from malaria through this mutation. Therefore, sickle cell anemia is common in areas affected by malaria (10).

1.4 Genetic basis of cancer

In all cases, cancer is caused by mutations in genes controlling the growth of cells and mitosis. Genes coding for proteins that are responsible for cell adhesion, growth and division are called proto-oncogenes. If a mutation occurs in these genes, they become oncogenes, which can cause cancer. On the other hand, the opponent of oncogenes - the tumor suppressor genes, inactivate oncogenes. Therefore, a mutation in tumor suppressor genes, that results in activation of oncogenes, can also be a cause of cancer development (15).

Fortunately, only a very small number of mutations are causing cancer. Most mutated cells die, because of a lack of power to survive compared to non-mutated cells and of these small number of mutated cells, again, only a few are progressing to cancer, because of still functioning feedback mechanism to control cell growth (15).

Furthermore, our immune system, which is also triggered by atypical proteins, can recognize mutating cells and destroy them. This is the reason for an elevated risk of the development of cancer in immunosuppressed patients. Finally, usually many different mutations in growth-promoting genes are needed for the development of cancer (15).

Nothing less than bad odds during the precise process of replication and repairing mechanism in the production of countless cells each year can be the reason for cancer.
Nevertheless, certain factors can increase the risk of occurrence of cancer:

- Aging,
- Ionizing radiation,
- Chemical substances (e.g. smoke),
- Physical irritants (e.g. heat),
- Hereditary tendencies (e.g. Lynch-Syndrome) and
- Viruses (e.g. HPV) (15).

1.5 Next-Generation Sequencing

The traditional method of sequencing, which is based on HLPC, that was developed by Sanger and therefore also called Sanger sequencing, is a very expensive method and had a small read-out, which made sequencing of only a small number of genes possible at a time. Nowadays, Sanger- sequencing is mostly replaced by Next-Generation Sequencing (NGS) technologies. These techniques have much higher throughput and are less expensive compared to standard sequencing. For instance, the human genome project, which took over a decade by Sanger sequencing, would now be possible in just one day (16).

NGS can roughly be divided into panel-sequencing and whole exome or genome sequencing. In whole-genome sequencing, the complete genome is sequenced and compared to the standard genome GRCh38. This also enables the inclusion of promotor and regulatory regions of the genome. This type of sequencing is a great approach for discovering uncommon or new mutations throughout the genome (16).

The sequencing of cDNA fragments, that were created from RNA via reverse transcriptase is called transcriptome-sequencing. The expression of RNA and variations in splicing alterations can be resolved in this way (16).

1.6 Bioinformatics

The discipline bioinformatics is a combination of sciences of biology, physics, mathematics and computation. It is a growing field in natural sciences for the management, analyzation and interpretation of huge amounts of data. The HGP and the following whole genome sequencing of many other species set a base for the application of bioinformatics. The two main instruments used by bioinformaticians are computer software programs, that are needed for an easy analyzation of data and the world wide web, to provide uncomplicated access to data (17).

1.6.1 Online databases

Because the quantity of biological data is rapidly increasing, online databases working with quick processing of information, a user-friendly scheme and algorithm software programs are required for the administration of these data in the bioinformatic analysis. There are different databases online, because of different sources of information. These databases can also be used by scientists and practitioners to check DNA sequences on the responsibility for causing disease, for example (18).

Bioinformatics is not only used on a genome level. In functional genomics, it also assesses resulting proteins and their function on the proteome and transcriptome level.
In the clinical setting, these databases could play a big role in the treatment of diseases caused by genetic mutations (18).

VarSome.com is one example of an online meta-database for the collection of biological information. It allows the exploration, assembling and investigation of the effects of human genetic variations. Experts from all over the world are sharing their knowledge of human genomics on one, easily accessible website. More than 30 databases are linked by VarSome. Over 33 billion data points express more than 500 million variants. A certain algorithm has been introduced to handle this large amount of data. Each variant has a certain location on the genome, similar variations are spotted, the type, such as frameshift, insertion-deletion, etc. is

determined, as well as the consequences of coding. VarSome can be used similarly to a browser. The matter in question can be searched for in different ways:
"gene name, transcript symbol, genomic location, variant ID, HGVS nomenclature or a single line from VCF files (19,20)". If there is a question about a gene or a transcript, the formal name of the gene, a description of the effect of the protein and possible disease in an association are listed.

The ACMG guideline classifies the questioned variant into one of five different groups:

- Pathogenic
- Likely pathogenic
- Variant of uncertain significance (VUS)
- Likely benign
- Benign (21)

Also, data will be presented concerning the frequency in the population. Pathogenicity prediction scores are determined by different *in silico* prediction tools. VarSome uses 133 datasets of *in silico* prediction tools (19).

1.6.2 *In silico* prediction tools

If only one nucleotide is altered in a DNA sequence it is called a single-nucleotide polymorphism (SNP). SNPs are responsible for the diversity in humans but can also contribute to the development of different diseases, such as cancer, diabetes or neurodegenerative diseases. There are many types of SNPs, but nonsynonymous single-nucleotide polymorphisms (nsSNPs) can change the structure, biochemistry and function of proteins regarding "folding characteristics, charge distribution, stability, dynamics, and interactions with other proteins (22)."

To predict easily and inexpensively the degree of consequences in the structure or function of a protein, biologists and scientists are using *in silico* prediction tools, working with computational algorithms (22).

In silico databases use mainly three approaches to determine a result for prediction. They are based on:

- Conservation/homology
- Protein structure and function
- Machine learning approaches (23)

2. OBJECTIVES

Firstly, the major aim of this work was to identify and evaluate the most important *in silico* databases, that are commonly used by VarSome, regarding basic principles of function, scores and thresholds. Furthermore, it was intended to assign the web link and PMID (PubMed identifier) number to each *in silico* database.

Secondly, the minor part of this work was to investigate in a pilot study, how well the classifications of yet unidentified genomic alterations in oncological samples, predicted by *in silico* databases, are congruent with the classifications determined by the molecular oncology of the department of pathology at the medical center in Coburg. The classifications of genomic alterations made by the department of pathology were established from careful literature-based single variant analysis.

3. MATERIALS AND METHODS

a. First (major) part: Identification and evaluation of *in silico* databases

The first (major) part of this work intended to extend and validate the analytical tools for variant classification used in the molecular oncology of the department of pathology at the medical center Coburg.

For this reason, VUS (variants of unknown significance), collected from the dataset of comp arrays from the year 2021 (n=92), were entered into VarSome. VarSome was chosen as the meta-database, because of routine use by pathologists at the medical center Coburg. The *in silico* tools, that appeared most often during the predictions of VUS, were identified as the most common in silico databases used by VarSome.

The evaluation of *in silico* tools included for each one:

- Basic principles of function
- Classification output (scores, thresholds)
- Web-link
- PMID number

This information was collected in the form of two tables for a better overview.

b. Second (minor) part: Pilot study

The second (minor) part of this study was conducted in form of a pilot study. It concentrated on the grade of congruency between classifications of VUS in oncological samples predicted by *in silico* databases and classifications determined via careful literature based single variant analysis by the molecular oncology of the department of pathology at the medical center in Coburg.

Therefore 92 genetic alterations taken from comp-arrays, that were collected from the year 2021, were used in this study. These were classified via careful literature based single variant analysis by the department of pathology according to the ACMG guidelines into benign, likely benign, VUS, likely pathogenic or pathogenic. The five categories were shortened to three categories for better comparison with predictions of *in silico* tools in:

- Benign (benign + likely benign)
- VUS
- Pathogenic (pathogenic + likely pathogenic).

Each genetic alteration was entered into VarSome, showing a prediction, derived from the average of multiple predictions made by different *in silico* databases. In this study, the most

common *in silico* databases used by VarSome were selected for the analyzation of the classifications made by the department of pathology.

Only genetic alterations, collected from comp-arrays, that were collected during the year 2021, were included in this study. Also, only variants, that could get an unambiguous predictive result were included.

Genetic alterations without a prediction result in every *in silico* prediction tool were excluded from this study. 58 genetic alterations remained for the analysis.

Statistical analysis

The grade of congruency was evaluated based on the comparison of the predicted classification made by each *in silico* database (benign/pathogenic) with the actual classification determined by the department of pathology via careful literature based single variant analysis (26 benign, 32 pathogenic).

After entering the 58 genetic alterations into VarSome, the predicted output from the *in silico* databases was used to create a binary confusion matrix (Figure 4).

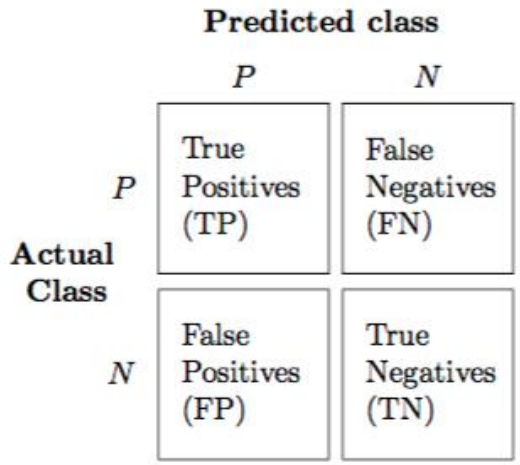

Figure 4. Confusion matrix

Source: http://rasbt.github.io/mlxtend/user_guide/evaluate/confusion_matrix/

Pathogenic predictions were encoded by the number one, whereas benign predictions received the coding number zero. These results were subdivided into 4 categories: true positive, true negative, false positive and false negative.

Logistic regression was used for the statistical analysis provided by the statistic program JASP version 0.16.3 (University of Amsterdam), to determine for each *in silico* tool:

- Accuracy [TP + TN / (TP + TN + FP + FN)] x 100
- Matthews Correlation Coefficient (MCC)
  (TP x TN) – (FP x FN) / √ (TP+FP)(TP+FN)(TN+FP)(TN+FN)

By comparing the actual classifications made by the pathology department with the predicted classifications from the *in silico* databases, the accuracy for every *in silico* database was calculated.

In contrast to accuracy, the MCC takes all four categories into account (true positive, true negative, false positive and false negative. Only if the *in silico* database reached good results in every category, the MCC will show high scores. The MCC score ranges from +1 (always correct prediction) to -1 (always false prediction), a score of 0 resembles a random prognostication. A MCC score of 0.5 or higher was set as the threshold for acceptability, which is equivalent to an accuracy of >75% (24).

4. RESULTS

a. First (major) part: Identification and evaluation of in silico databases

After entering 92 genetic alterations, collected from the dataset of molecular oncology of the department of pathology at the medical center Coburg, into VarSome, 20 *in silico* databases were identified as the most important prediction tools commonly used by VarSome. These are BayesDel addAF, BayesDel noAF, DEOGEN2, MetaLR, MetaRNN, MetaSVM, REVEL, EIGEN, EIGEN PC, FATHMM, FATHMM-MKL, FATHMM-XF, LIST-S2, LRT, Mutation Assessor, MutationTaster2, PrimateAl, PROVEAN, SIFT and SIFT4G (Table 1 and Table 2).

In Table 1, the basic mechanism of function was summarized for each *in silico* database. The *in silico* databases were subdivided into meta scores and individual prediction tools according to their basic mechanism of function (Table 1). Meta scores combine the results of different individual prediction tools to obtain a prediction result.

Table 1. Basic principles of *in silico* databases

***In silico* databases**	**Basic principles of function**
Meta scores	
BayesDel addAF **BayesDel add allele frequency** **BayesDel noAF** **BayesDel no allele frequency**	"met-score that combines deleteriousness predictors in the naïve Bayesian approach and uses ClinVar (Landrum et al., 2014) variants as a standard to determine the cutoff value. For this predictor, the set that integrates maximum and minor allele frequencies across populations (addAF) presents superior performance to that without allele frequencies (noAF) (Feng, 2017) (24)"
DEOGEN2	"protein sequence-based predictor that utilizes evolutionary information as well as contextual information, such as the relevance of the gene containing the variant or the interactions of the encoded protein (25)"
MetaLR	"MetaSVM and MetaLR are two ensemble scores based on Support Vector Machine (SVM) and Logistic Regression (LR), respectively. Both methods integrate the information of 11 non-ensemble predictors (PolyPhen-2, SIFT, MutationTaster, Mutation Assessor, FATHMM, LRT, PANTHER, PhD-SNP, SNAP, SNPs&GO and MutPred), three conservation scores (GERP++, SiPhy and PhyloP) and four ensemble scores (CADD, PON-P, KGGSeq and CONDEL (26)"
MetaRNN **Meta recurrent neural network**	"recurrent neural network (RNN) based ensemble prediction score, which incorporated 16 scores (SIFT, Polyphen2_HDIV, Polyphen2_HVAR, MutationAssessor, PROVEAN, VEST4, M-CAP, REVEL, MutPred, MVP, PrimateAI, DEOGEN2, CADD, fathmm-XF, Eigen and GenoCanyon), 8 conservation scores (GERP, phyloP100way_vertebrate, phyloP30way_mammalian, phyloP17way_primate, phastCons100way_vertebrate, phastCons30way_mammalian, phastCons17way_primate and SiPhy), and allele frequency information from the 1000 Genomes Project (1000GP), ExAC, and gnomAD (27)"

MetaSVM **Meta-analytic support vector machine**	"MetaSVM and MetaLR are two ensemble scores based on Support Vector Machine (SVM) and Logistic Regression (LR), respectively. Both methods integrate the information of 11 non-ensemble predictors (PolyPhen-2, SIFT, MutationTaster, Mutation Assessor, FATHMM, LRT, PANTHER, PhD-SNP, SNAP, SNPs&GO and MutPred), three conservation scores (GERP++, SiPhy and PhyloP) and four ensemble scores (CADD, PON-P, KGGSeq and CONDEL (26)"
REVEL **Rare exom variant ensemble learner**	"based on a combination of scores from 13 individual tools: MutPred, FATHMM v2.3, VEST 3.0, PolyPhen-2, SIFT, PROVEAN, MutationAssessor, MutationTaster, LRT, GERP++, SiPhy, phyloP, and phastCons. REVEL was trained using recently discovered pathogenic and rare neutral missense variants, excluding those previously used to train its constituent tools (28)"
Individual prediction tools	
EIGEN	EIGEN makes use of a variety of functional annotations in both coding and noncoding regions (such as made available by the ENCODE and Roadmap Epigenomics projects), and combines them into one single measure of functional importance. Eigen is an unsupervised approach, and, unlike most existing methods, is not based on any labelled training data. Eigen produces estimates of predictive accuracy for each functional annotation score, and subsequently uses these estimates of accuracy to derive the aggregate functional score for variants of interest as a weighted linear combination of individual annotations (29)"
EIGEN PC	"based on the eigendecomposition of the annotation covariance matrix, and using the lead eigenvector to weight the individual annotations (30)"
FATHMM **Functional analysis through hidden markov models**	"Our software and server is capable of predicting the functional effects of protein missense mutations by combining sequence conservation within hidden Markov models (HMMs), representing the alignment of homologous sequences and conserved protein domains, with "pathogenicity weights", representing the overall tolerance of the protein/domain to mutations. (31)"
FATHMM- MKL **Fathmm- multiple kernel learning**	"integrates functional annotation information from the ENCODE with nucleotide-based HMMs (32)"
FATHMM- XF **Fathmm with extended features**	"By using an extended set of feature groups and by exploring an expanded set of possible models, the new method yields even greater accuracy than its predecessor on independent test sets. Unlike FATHMM-MKL, FATHMM-XF models are build up on single-kernel datasets. The models may then learn interactions between data sources that help to boost its accuracy in all regions of the genome (33)"
LIST- S2	"first, it aligns (high Local identity Pair-wise Sequence Alignment, LPSA) the query sequence to all protein sequences in the UniProt Swiss-Prot/TrEMBL database and then it identifies the most relevant homologies based on their local identity to the query sequence around that position. And finally, it estimates the potential deleteriousness of mutations based on Taxonomy distance of species with variations to the query (34)"

LRT **Likelyhood ratio test**	"using a comparative genomics data set of 32 vertebrate species (35)"
Mutation Assessor	"conservation-based approach. It distinguishes between conservation patterns within aligned families (conservation score) and sub-families (specificity score) of homologs and so attempts to account for functional shifts between subfamilies of proteins (36)"
MutationTaster2	"uses regulatory features, degree of evolutionary conservation and splice site predictions as the input for a naïve Bayes classifier (37)"
PrimateAI	"trained on a dataset of ~380,000 common missense variants from humans and six non-human primate species, using a semi-supervised benign vs unlabeled training regimen. The input to the network is the amino acid sequence flanking the variant of interest and the orthologous sequence alignments in other species, without any additional human-engineered features. To incorporate information about protein structure, PrimateAI learns to predict secondary structure and solvent accessibility from amino acid sequence and includes these as sub-networks in the full model. The total size of the network, with protein structure included, is 36 layers of convolutions, consisting of roughly 400,000 trainable parameters (38)"
PROVEAN **Protein variation effect analyzer**	"based on the change, caused by a given variation, in the similarity of the query sequence to a set of its related protein sequences (39)"
SIFT **Sorting intolerant from tolerant**	"based on sequence homology and the physical properties of amino acids (40)"
SIFT4G **Sorting intolerant from tolerant for genomes**	"faster version of SIFT (41)"

In Table 2, web links, classification output in form of scores and threshold, as well as the PMID number were summarized for each *in silico* database.

The web link leads to the homepage of the corresponding *in silico* database. Data for prediction can also be directly entered at this site.

The classification output describes the specific score and threshold belonging to each *in silico* database for classifying predictions. Mostly 2 categories are defined by the threshold: likely benign/ benign/ tolerated/ neutral/ polymorphism and likely pathogenic/ deleterious/ damaging/ pathogenic/ disease causing. Exceptions are Mutation Assessor, which has 4 categories defined by thresholds: neutral, low, medium and high and MutationTaster2, which splits its results into "the prediction is true" or "MutationTaster2 comes to a different conclusion".

The PMID number directly leads to the research paper of the corresponding *in silico* database, published by the developer of the *in silico* tool.

Table 2. Web links, classification output and PMID number of *in silico* databases

***In silico* databases**	**Web-links**	**Classification output**	**PMID number**
Meta Scores			
BayesDel addAF **BayesDel add allele frequency** **BayesDel noAF** **BayesDel no allele frequency**	https://fengbj-laboratory.org/BayesDel/BayesDel.html	score: -1.29334 to 0.75731 threshold with MaxAF: <0.0692655: likely benign ≥0.0692655: likely pathogenic others: uncertain significance threshold without MaxAF: <-0.0570105: likely benign ≥-0.0570105: likely pathogenic others: uncertain significance	31484976
DEOGEN2	https://bio.tools/DEOGEN2	score: 0 to 1 threshold: <0.5: tolerated ≥0.5: deleterious	28498993
MetaLR	no results found	score: 0 to 1 threshold: <0.5 tolerated ≥0.5 damaging	
MetaRNN **Meta recurrent neural network**	http://www.liulab.science/metarnn.html	score: 0 to 1 threshold: <0.5 tolerated ≥0.5 damaging	https://doi.org/10.1101/2021.04.09.438706
MetaSVM **Meta-analytic support vector machine**	no results found	score: -2.0058 to 3.0399 threshold: <0: tolerated ≥0: damaging	28149325 https://doi.org/10.1101/805051
REVEL **Rare exom variant ensemble learner**	https://sites.google.com/site/revelgenomics/about?authuser=0	score: 0 to 1 threshold: <0.5: benign ≥0.5: pathogenic	27666373
Individual prediction tools			
EIGEN **EIGEN PC**	http://www.columbia.edu/~ii2135/eigen.html	score: threshold: the larger the score the more likely the variant has a damaging effect	26727659
FATHMM **Functional analysis through hidden markov models**	http://fathmm.biocompute.org.uk/	score: -16.13 to 10.64 threshold: >-1.5: tolerated ≤-1.5: damaging	23033316 23620363 24980617

FATHMM-MKL **Fathmm- multiple kernel learning**	http://fathmm.biocompute.org.uk/fathmmMKL.htm	score: 0 to 1 threshold: > 0.5: deleterious ≥ 0.7: pathogenic ≤ 0.5: neutral	25583119
FATHMM- XF **Fathmm with extended features**	http://fathmm.biocompute.org.uk/fathmm-xf/	score: 0 to 1 threshold: ≥0.5: deleterious <0.5: neutral/benign	28968714
LIST- S2	https://precomputed.list-s2.msl.ubc.ca/	score: 0 to 1 threshold: <0.85: benign ≥0.85: deleterious	32352516
LRT **Likelyhood ratio test**	No results found	score: 0 to 1 threshold: <0.001: deleterious ≥0.001: neutral	19602639
Mutation Assessor	http://mutationassessor.org/r3/	score: -5.135 to 6.49 threshold: <1: neutral (benign) ≥1: low (pathogenic) ≥2: medium (pathogenic) ≥3.5: high (pathogenic)	17976239
MutationTaster2	https://www.genecascade.org/MutationTaster2021/	score: 0 to 1 (probability that the prediction is true) threshold: 1: prediction is true <0.5: MT classifier comes to a different conclusion disease causing – polymorphism ('automatic' is added when the effect of the variant has already been clarified)	24681721 https://doi.org/10.1093/nar/gkab266
PrimateAI	https://github.com/Illumina/PrimateAI	score: 0 to 1 threshold: <0.8: tolerated ≥0.8: damaging	30038395
PROVEAN **Protein variation effect analyzer**	http://provean.jcvi.org/index.php	score: -14 to 14 threshold: ≤ predefined threshold: deleterious (-2) > predefined threshold: neutral (-2)	25851949 23056405

SIFT **Sorting intolerant from tolerant** **SIFT4G** **Sorting intolerant from tolerant for genomes**	https://sift.bii.a-star.edu.sg/ https://sift.bii.a-star.edu.sg/sift4g/	score: 0 to 1 threshold: ≤0.05: deleterious >0.05: tolerated	22689647 11337480 19561590 11875032 12824425 26633127 26633127

b. Second (minor) part: Pilot study

In the second (minor) part of the study, the congruency between classification determined via careful literature based single variant analysis and classifications made by *in silico* prediction tools was investigated through observing the accuracy and MCC of each *in silico* tool.

Therefore, 58 VUS, out of which 32 were classified as pathogenic and 26 classified as benign by the department of pathology via careful literature based single variant analysis, were analyzed by 20 *in silico* databases (Figure 5, Figure 6).

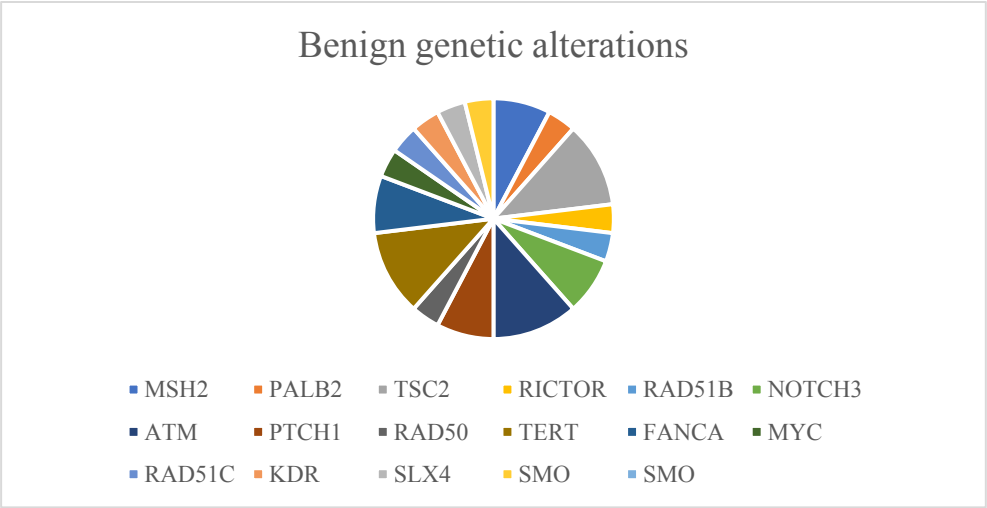

Figure 5. Benign genetic alterations classified by the department of pathology (n =26)

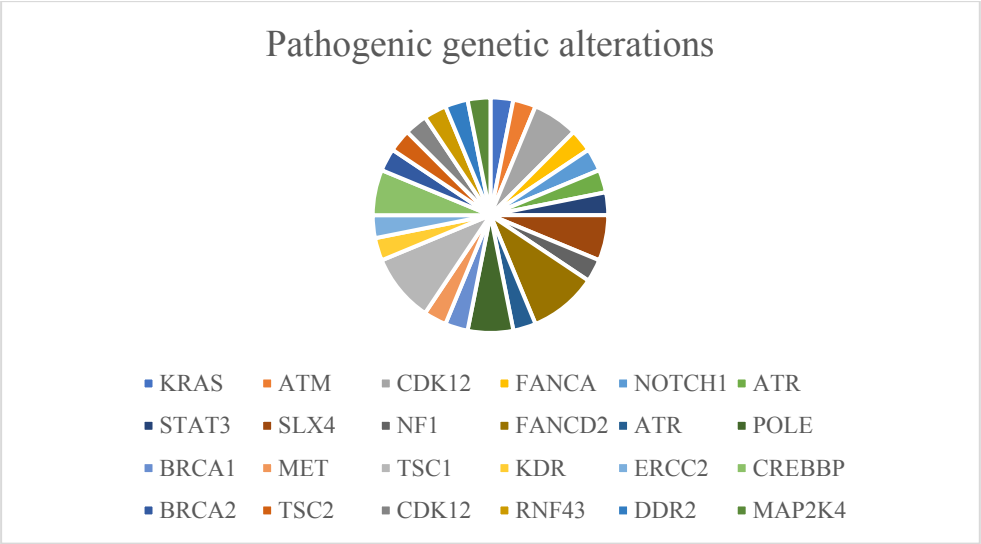

Figure 6. Pathogenic genetic alterations classified by the department of pathology (n = 32)

The results of the *in silico* databases regarding accuracy and MCC are shown in Table 3.

Table 3 Evaluation of predictions of *in silico* databases

***In silico* database**	**Accuracy (%)**	**MCC score**
DEOGEN	53,4	0,14
MetaLR	53,4	0,17
MetaRNN	60,3	0,39
MetaSVM	51,7	0,12
MutationTaster	58,6	0,15
REVEL	55,2	0,23
BayesDel addAF	65,5	0,46
BayesDel noAF	58,6	0,23
EIGEN	50,0	0,01
EIGEN PC	55,2	0,09
FATHMM	51,7	0,02
FATHMM-MKL	55,2	0,06
FATHMM-XF	51,7	0,02
LIST-S2	56,9	0,13
LRT	55,2	0,08
Mutation Assessor	51,7	0,03
PrimateAI	51,7	0,19
Provean	53,4	0,10
Sift	55,2	0,14
Sift4G	56,9	0,20

The highest value of each category is highlighted in green
The lowest value of each category is highlighted in red

The results regarding the accuracy of *in silico* databases ranged from 50,0%% to 65,5%. The best result was shown by BayesDel addAF with 65,5%. EIGEN showed the lowest result with an accuracy of 50% (Figure 7).

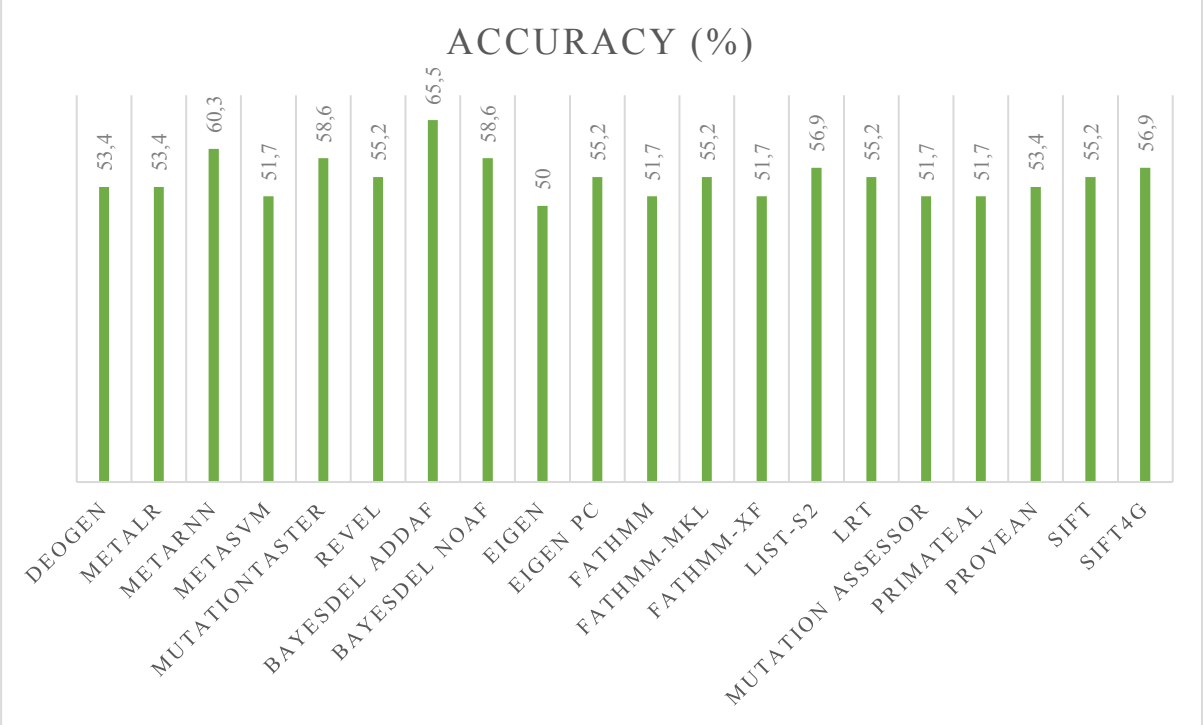

Figure 7. Accuracy of *in silico* databases

Based on the MCC score, BayesDel addAF scored highest with 0.46. The lowest result was shown by EIGEN with 0.01. Every *in silico* database stayed below the cutoff value of 0,5 for acceptable performance (Figure 8).

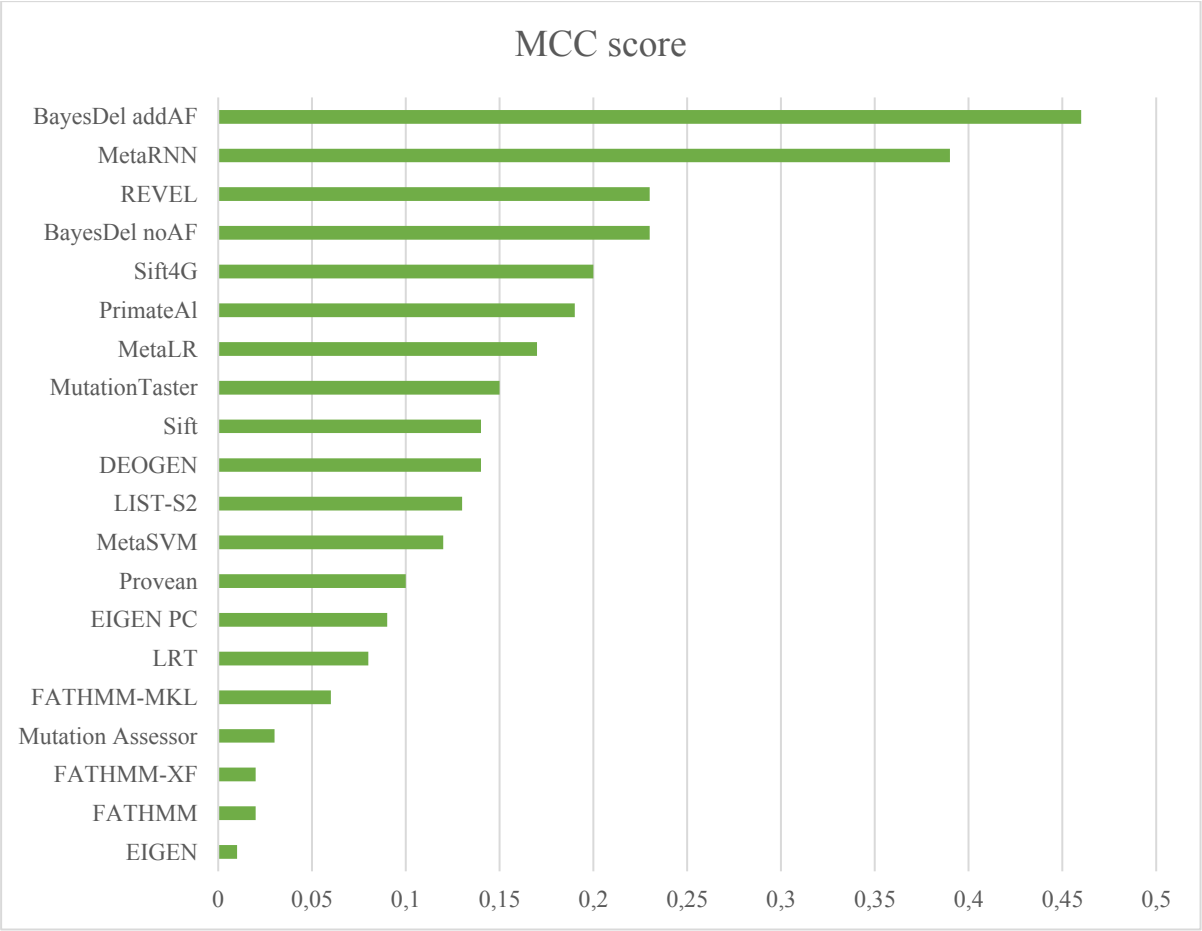

Figure 8. MCC score of *in silico* databases

5. DISCUSSION

Bioinformatics is becoming a growing field in medicine, especially important for personalized medicine. In this context, *in silico* prediction tools were developed to predict the effects of genetic alterations (21). The possibility of predicting the effects of unknown genetic alterations could also be of great use for the molecular oncology of the department of pathology at the medical center in Coburg, regarding the classification of genomic alterations in oncological samples.

The first (major) part of this work was concentrated on the identification and evaluation of the most important *in silico* tools, that are commonly used by VarSome. VarSome is the meta-database, that is routinely used by pathologists at the medical center Coburg.

20 *in silico* tools were identified, that were evaluated regarding the basic principle of function, classification output, including scores and thresholds, as well as web-link and PMID number. This information (summarized in Table 1 and Table 2) enables pathologists from the molecular oncology of the department of pathology at the medical center Coburg to obtain a greater and deeper understanding of *in silico* databases and by this, improve the work with these prediction tools. Also, a possibility for quick further research is given by following the web link and PMID number for each *in silico* database.

The evaluation of additional *in silico* databases can be investigated in further research.

The second (minor) part of this work observed the congruity between classifications of genomic alterations in oncological samples determined by the department of pathology via careful literature-based single variant analysis and the classifications predicted by *in silico* tools. 32 VUS, that were classified as pathogenic and 26 VUS, that were classified as benign, by the department of pathology via careful literature-based single variant analysis, were used in this study. The evaluation of the congruity between the classifications made by the pathology department and the *in silico* databases was accomplished by studying the accuracy and MCC of each prediction tool.

The results of this study showed that none of the *in silico* databases reached the acceptable value of 0,5 in the MCC score (equivalent to >75% accuracy). This was also reflected in the accuracy, that only showed a range of 50% to 65,5%. These results contrast with a study conducted by Ernst *et al.* In this study an accuracy of 92% could be reached (42). This discrepancy can be explained by different limitations in our study.

Firstly, the system of classifying VUS by the department of pathology could be biased at different levels. This could lead to prediction results of *in silico* databases, that are falsely

classified as false negative/positive, which also has a negative influence on the overall performance of *in silico* tools. Secondly, only VUS were included in our study. However, *in silico* databases can reach better results, when they are tested on a higher collective, without the exclusion of truly positive/negative variants. This was also shown in the study conducted by Leong *et al*., that included in total 283 pathogenic and 29 benign gene variants in their study and could reach accuracies of >90% (24).

The findings of our study are relevant for gaining experience and knowledge about the use of *in silico* databases at the molecular oncology of the department of pathology, regarding the classification of genomic alterations in oncological samples. As a consequence of the non-congruent results of our study, the analyzing strategy of variant classification by *in silico* tools should be rethought. *In silico* tools cannot replace the literature-based single variant analysis of genomic alterations, yet.

Improvement can be achieved by checking the way of classifying genomic alterations in the department of pathology. Moreover, a larger dataset of genomic alterations, in which known benign/pathogenic variants are included, should be used in further studies.

-

-

6. CONCLUSION

In conclusion, the results of the first (major) part of this work give the opportunity, by identifying and evaluating the most important *in silico* databases commonly used by VarSome, to extend the interpretation and validation of *in silico* databases, that are frequently used at the molecular oncology of the department of pathology at the medical center Coburg.

The results of the second (minor) part, that investigated the congruency between classifications established through careful literature-based single variant analysis by the department of pathology and the classifications made by *in silico* predictions tools are indicative, but low in congruency. The results of the pilot study have contributed to get a feeling for the reliability of in silico databases. However, the literature-based single variant analysis cannot be replaced by predictions of *in silico* databases, yet.

7. REFERENCES

1. Nature education [Internet]. Scitable by nature education: Nature Publishing Group; 2014. Definition Genome; [cited 2022 August 25]. Available from: https://www.nature.com/scitable/definition/genome-43/
2. Nature education [Internet]. Scitable by nature education: Nature Publishing Group; 2014. Definition Human Genome Project; [cited 2022 August 25]. Available from: https://www.nature.com/scitable/definition/human-genome-project-112/
3. Collins F, Fink L. The Human Genome Project. Alcohol Health Res World. 1995;19:190-5.
4. Nature [Internet]. Nature article: Sara Reardon; 2021. A complete human genome is close: how scientists filled in the gaps; [cited 2022 August 25]. Available from: https://www.nature.com/articles/d41586-021-01506-w
5. Nature education [Internet]. Scitable by nature education: Nature Publishing Group; 2014. Concept protein function; [cited 2022 August 25]. Available from: https://www.nature.com/scitable/topicpage/protein-function-14123348/
6. ER services [Internet]. Biology for majors 1: Shelli Carter and Lumen Learning; 2016. Proteins – Describe the structure and function; [cited 2022 August 25]. Available from: https://courses.lumenlearning.com/suny-wmopen-biology1/chapter/proteins/#:~:text=Proteins%20are%20polymers%20of%20amino,and%20a%20variable%20R%20group
7. Nature education [Internet]. Scitable by nature education: Nature Publishing Group; 2014. Concept protein structure; [cited 2022 August 25]. Available from: https://www.nature.com/scitable/topicpage/protein-structure-14122136/
8. Thieme via medici [Internet]. Via medici – informieren: Redaktion via medici; 2008. Räumliche Anordnung von Proteinen; [2022 August 25]. Available from: https://m.thieme.de/viamedici/vorklinik-faecher-biochemie-1511/a/raeumliche-anordnung-von-proteinen-3845.htm
9. NIH National cancer institute [Internet]. NCI Dictionaries: 2022. Mutation; [cited 2022 August 25]. Available from: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/mutation
10. Studyflix [Internet]. Genetik: 2022. Mutation; [cited 2022 August 25]. Available from: https://studyflix.de/biologie/mutation-2582
11. Studyflix [Internet]. Genetik: 2022. Genommutation; [cited 2022 August 25]. Available from: https://studyflix.de/biologie/genommutation-2572

12. Studyflix [Internet]. Genetik: 2022. Chromosomenmutation; [cited 2022 August 25]. Available from: https://studyflix.de/biologie/chromosomenmutation-2571
13. Studyflix [Internet]. Genetik: 2022. Genmutation; [cited 2022 August 25]. Available from: https://studyflix.de/biologie/genmutation-2484
14. López-Urrutia E, Salazar-Rojas V, Brito-Elías L, Coca-González M, Silva-García J, Sánchez-Marín D et al. BRCA mutations: is everything said? Breast Cancer Res Treat. 2019;173:49-54.
15. Guyton AC, Hall JE. Cancer.In: Guyton and Hall Textbook of Medical Physiology. 13th ed. Philadelphia: Elsevier; 2016. p. 41-43.
16. Guan Y, Li G, Wang R, Yi Y, Yang L, Jiang D et al. Application of next-generation sequencing in clinical oncology to advance personalized treatment of cancer. Chin J Cancer. 2012;31:463-70.
17. Moore J. Bioinformatics. J Cell Physiol. 2007;213:365-9.
18. Bayat A. Science, medicine, and the future: Bioinformatics. BMJ. 2002;324:1018-22.
19. Kopanos C, Tsiolkas V, Kouris A, Chapple C, Albarca Aguilera M, Meyer R et al. VarSome: the human genomic variant search engine. Bioinformatics. 2019;35:1978-80.
20. den Dunnen J, Dalgleish R, Maglott D, Hart R, Greenblatt M, McGowan-Jordan J et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Hum Mutat. 2016;37:564-9.
21. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J et al. ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17:405-24.
22. Yazar M, Özbek P. In Silico Tools and Approaches for the Prediction of Functional and Structural Effects of Single-Nucleotide Polymorphisms on Proteins: An Expert Review. OMICS. 2021;25:23-37.
23. Leong I, Stuckey A, Lai D, Skinner J, Love D. Assessment of the predictive accuracy of five in silico prediction tools, alone or in combination, and two metaservers to classify long QT syndrome gene mutations. BMC Med Genet. 2015;16:34.
24. Borges P, Pasqualim G, Matte U. Which Is the Best In Silico Program for the Missense Variations in IDUA Gene? A Comparison of 33 Programs Plus a Conservation Score

and Evaluation of 586 Missense Variants. Front Mol Biosci. 2021. doi: 10.3389/fmolb.2021.752797.

25. Ancien F, Pucci F, Vranken W, Rooman M. MutaFrame - an interpretative visualization framework for deleteriousness prediction of missense variants in the human exome. Bioinformatics. 2021;38:265–6.
26. Alarcon J, Enriquez J, Sanchez-Cabo F. Frequency Conservation Score (FCS): the power of conservation and allele frequency for variant pathogenic prediction. Biorxiv. 2019. doi: 10.1101/805051
27. Liu lab science. MetaRNN [Internet]. Biorxiv: Chang L, Degui Z, Kai W, Xiaoming L; [21.10.2021]. Available from: http://www.liulab.science/metarnn.html
28. Ioannidis N, Rothstein J, Pejaver V, Middha S, McDonnell S, Baheti S et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016;99:877-85.
29. EIGEN [Internet]. Columbia University: IULIANA IONITA-LAZA I, MCCALLUM K,[1]XU B, BUXBAUM J; 2015. About EIGEN; [cited 2022 August 2025] Available from: http://www.columbia.edu/~ii2135/information_eigen.html
30. Ionita-Laza I, McCallum K, Xu B, Buxbaum J. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet. 2016;48:214-20.
31. Fathmm [Internet]. University of Bristol: Shihab H, Gough J, Cooper D, Stenson P, Barker G, Edwards K, Day I, Gaunt T; 2013. About fathmm; [cited 2022 August 25]. Available from: http://fathmm.biocompute.org.uk/about.html#coding
32. Shihab H, Rogers M, Gough J, Mort M, Cooper D, Day I et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics. 2015;31:1536-43.
33. Rogers M, Shihab H, Mort M, Cooper D, Gaunt T, Campbell C. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics. 2018;34:511-13.
34. LIST-S2 [Internet]. Nucleic acids research: Malhis N, Jacobson M, Jones S, Gsponer J; 2020. LIST-S2 About; [cited 2022 August 25]. Available from: https://list-s2.msl.ubc.ca/about
35. Chun S, Fay J. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19:1553-61.

36. Gnad F, Baucom A, Mukhyala K, Manning G, Zhang Z. Assessment of computational methods for predicting the effects of missense mutations in human cancers. BMC Genomics. 2013;14 Suppl 3:S7.
37. Ernst C, Hahnen E, Engel C, Nothnagel M, Weber J, Schmutzler R et al. Performance of in silico prediction tools for the classification of rare BRCA1/2 missense variants in clinical diagnostics. BMC Med Genomics. 2018;11:35.
38. Illumina [Internet]. Nature Genetics: Sundaram L, Gao H, Padigepati S, McRae J, Li Y, Kosmicki J et al; 2018. Primate Al; [cited 2022 August 25]. Available from: https://github.com/Illumina/PrimateAI/blob/master/README.md
39. Yongwook C. A fast computation of pairwise sequence alignment scores between a protein and a set of single-locus variants of another protein. BCB. 2012;12:414-17.
40. Sorting Intolerant from Tolerant [Internet]. NIH: Sim N, Kumar P, Hu J, Henikoff S, Schneider G, Ng P; 2012. Sorting intolerant from tolerant home; [cited 2022 August 25]. Available from: https://sift.bii.a-star.edu.sg/
41. Vaser R, Adusumalli S, Leng S, Sikic M, Ng P. SIFT missense predictions for genomes. Nat Protoc. 2016;11:1-9.
42. Ernst C, Hahnen E, Engel C, Nothnagel M, Weber J, Schmutzler R et al. Performance of in silico prediction tools for the classification of rare BRCA1/2 missense variants in clinical diagnostics. BMC Med Genomics. 2018;11:35.

8. SUMMARY

Objectives: The aim of the first (major) part of this study was to identify and evaluate the most important *in silico* databases, that are commonly used by VarSome, regarding basic principles of function, classification output (scores and thresholds), web link and PMID number. The second (minor) aim of this study was to investigate in a pilot study, how well the classifications of yet unidentified genomic alterations in oncological samples, predicted by *in silico* databases, are congruent with the classifications determined via careful literature-based single variant analysis by the molecular oncology of the department of pathology at the medical center in Coburg.

Materials and methods: In the first (major) part of this study, 92 VUS (variants of unknown significance), were entered into VarSome and the most important ones were identified. The results were clearly presented in the form of two tables. The second (minor) part of this work was conducted in form of a pilot study. 58 genetic variants were entered into VarSome for prediction. The predictive results (benign/pathogenic) were compared to the classifications determined by the department of pathology. Accuracy and MCC score were determined for each *in silico* database. A MCC score of 0.5 or higher was set as the threshold for acceptability, which is equivalent to an accuracy of >75%.

Results: 20 *in silico* databases were identified as the most important prediction tools commonly used by VarSome (BayesDel addAF, BayesDel noAF, DEOGEN2, MetaLR, MetaRNN, MetaSVM, REVEL, EIGEN PC, FATHMM, FATHMM-MKL, FATHMM-XF, LIST-S2, LRT, Mutation Assessor, MutationTaster2, PrimateAl, PROVEAN, SIFT, SIFT 4G). The classification output (scores, thresholds), assigned to each tool, gives the possibility for a deeper understanding of predictive results. Further research can be quickly achieved by following the web link or PMID number determined for each *in silico* database. In the pilot study, the testing of congruency was indicative, but remained low in congruency. This was reflected in an MCC score, that remained overall below the acceptability value of 0.5.

Conclusion: The results of the first (major) part of this work enable pathologists to extend the knowledge about interpreting and validating *in silico* databases, that are frequently used at the molecular oncology of the department of pathology at the medical center Coburg. The results of the second (minor) part, the pilot study, have contributed to get a feeling for the reliability of in silico databases. However, careful literature-based single variant analysis cannot be replaced by predictions of *in silico* databases, yet.

9. CROATIAN SUMMARY

PROCJENA BAZA PODATAKA *IN SILICO* ZA KLASIFIKACIJU GENOMSKIH PROMJENA U ONKOLOŠKIM UZORCIMA

Ciljevi: Cilj prvog (glavnog) dijela ove studije bio je identificirati i evaluirati najvažnije in silico baze podataka, koje VarSome obično koristi, u vezi s osnovnim načelima funkcioniranja, izlazom klasifikacije (rezultati i pragovi), web poveznicom i PMID brojem . Drugi (manji) cilj ove studije bio je u pilot studiji istražiti koliko su klasifikacije još neidentificiranih genomskih promjena u onkološkim uzorcima, predviđene bazama podataka in silico, u skladu s klasifikacijama utvrđenim pažljivom analizom jedne varijante temeljenom na literaturi. od strane molekularne onkologije odjela patologije u medicinskom centru u Coburgu.

Materijali i metode: U prvom (velikom) dijelu ovog istraživanja u VarSome su unesene 92 VUS (varijante nepoznatog značaja) te su identificirane one najvažnije. Rezultati su pregledno prikazani u obliku dvije tablice. Drugi (manji) dio ovog rada proveden je u obliku pilot studije. 58 genetskih varijanti uneseno je u VarSome za predviđanje. Prediktivni rezultati (benigni/patogeni) uspoređeni su s klasifikacijama koje je odredio odjel patologije. Točnost i MCC rezultat određeni su za svaku in silico bazu podataka. MCC rezultat od 0,5 ili viši postavljen je kao prag prihvatljivosti, što je ekvivalentno točnosti od >75%.

Rezultati: 20 baza podataka in silico identificirano je kao najvažniji alati za predviđanje koje VarSome obično koristi (BayesDel addAF, BayesDel noAF, DEOGEN2, MetalLR, MetaRNN, MetaSVM, REVEL, EIGEN PC, FATHMM, FATHMM-MKL, FATHMM-XF, LIST-S2, LRT, Mutation Assessor, MutationTaster2, PrimateAl, PROVEAN, SIFT, SIFT 4G). Izlaz klasifikacije (rezultati, pragovi), dodijeljen svakom alatu, daje mogućnost za dublje razumijevanje prediktivnih rezultata. Daljnje istraživanje može se brzo postići praćenjem web poveznice ili PMID broja određenog za svaku in silico bazu podataka. U pilot studiji, ispitivanje kongruencije bilo je indikativno, ali je ostalo niske kongruencije. To se odrazilo na MCC ocjenu, koja je u cjelini ostala ispod vrijednosti prihvatljivosti od 0,5.

Zaključci: Rezultati prvog (velikog) dijela ovog rada omogućuju patolozima da prošire znanje o interpretaciji i validaciji in silico baza podataka, koje se često koriste na molekularnoj onkologiji odjela patologije medicinskog centra Coburg. Rezultati drugog (manjeg) dijela, pilot studije, pridonijeli su stjecanju osjećaja pouzdanosti in silico baza podataka. Međutim, pažljiva

analiza pojedinačnih varijanti temeljena na literaturi još se ne može zamijeniti predviđanjima baza podataka in silico.

10. CURRICULUM VITAE

Personal Information

Name: Véronique Agnes Amann

Date and place of birth: 14.07.1997, Ottweiler, Germany

Nationality: German

Address:

E-Mail: veronique1479@hotmail.de

Education

Since 10/2016: Medical Studies in English at the University of Split, School of Medicine

06/2016: Abitur at Heinrich-Böll-Gymnasium Saalfeld

Other activities

I am interested in different sports and music.