Effects of genetic variation on the structure of biological macromolecules

Jingxuan Kang[1], Siyu Wei[1], Zhe Jia[1], Yingnan Ma[1], Haiyan Chen[1], Chen Sun[1], Jing Xu[1], Junxian Tao[1], Yu Dong[1], Wenhua Lv[1], Hongsheng Tian[1], Xuying Guo[1], Shuo Bi[1], Chen Zhang[1], Yongshuai Jiang[1], Hongchao Lv[1], and Mingming Zhang[1]

[1]Harbin Medical University

June 1, 2023

Abstract

Changes in the structure of biological macromolecules, such as RNA and protein, have an important impact on biological functions, and are even important determinants of disease pathogenesis and treatment. Some genetic variations, including copy number variation, single nucleotide variation, and so on, can lead to changes in biological function and increased susceptibility to certain diseases by changing the structure of biological macromolecules. Here, we reviewed the progress of research about the effects of genetic variation on the structure of macromolecules including RNAs and proteins, several typical methods and common tools, and the effect on several diseases. An online resource (http://www.onethird-lab.com/gems/) to support convenient retrieval of common tools is also built. Finally, the challenges and future development of effect prediction were discussed.

Effects ofgenetic variation on the structure of biological macromolecules

Jingxuan Kang[1,2,+], Siyu Wei[1,2,+], Zhe Jia[1,2,+], Yingnan Ma[1,2,+], Haiyan Chen[1,2,+], Chen Sun[1,2], Jing Xu[1,2], Junxian Tao[1,2], Yu Dong[1,2], Wenhua Lv[1], Hongsheng Tian[1], Xuying Guo[1], Shuo Bi[1], Chen Zhang[1],

Yongshuai Jiang[1,2,*], Hongchao Lv[1,2,*], Mingming Zhang[1,2,*]

[1]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China.

[2]The EWAS Project, China.

[+]These authors contributed equally to this work.

[*]Correspondence:

Mingming Zhang, College of Bioinformatics Science and Technology, Harbin Medical University, 194 Xuefu Road, Nangang District, Harbin, Heilongjiang Province, China. E-mail: zhangmingming@hrbmu.edu.cn

Hongchao Lv, College of Bioinformatics Science and Technology, Harbin Medical University, 194 Xuefu Road, Nangang District, Harbin, Heilongjiang Province, China. E-mail: lvhongchao@gmail.com;

Yongshuai Jiang, College of Bioinformatics Science and Technology, Harbin Medical University, 194 Xuefu Road, Nangang District, Harbin, Heilongjiang Province, China. E-mail: jiangyongshuai@gmail.com;

Abstract

Changes in the structure of biological macromolecules, such as RNA and protein, have an important impact on biological functions, and are even important determinants of disease pathogenesis and treatment. Some

genetic variations, including copy number variation, single nucleotide variation, and so on, can lead to changes in biological function and increased susceptibility to certain diseases by changing the structure of biological macromolecules. Here, we reviewed the progress of research about the effects of genetic variation on the structure of macromolecules including RNAs and proteins, several typical methods and common tools, and the effect on several diseases. An online resource (*http://www.onethird-lab.com/gems/*) to support convenient retrieval of common tools is also built. Finally, the challenges and future development of effect prediction were discussed.

Keywords: Genetic variation; single nucleotide variation; macromolecular structure; RNA; protein.

Introduction

There are many forms of genetic variation, large to structural fragment insertion/deletion, and copy number variation (CNV), small to short insertion/deletion, and single nucleotide variation (SNV), the effects of which are complex. Genetic variation and lead to changes in gene function or affect the expression quantitative trait loci (eQTL) and cause abnormal gene expression. For biological macromolecules, genetic variation may change the structural sequence directly, and cause abnormalities in the biological process. For example, Ennis et al.[1] found that T/C (ss71651738) single nucleotide polymorphisms (SNPs) located at the origin of replication may lead to changes in replication forks and repeated expansion. The resulting repetitive instability is thought to be an important cause of Fragile X syndrome (FXS) [2], which is caused by *FMR1*silencing due to repeated expansion of base CGG [3].

RNA molecules fold into complex structures due to intramolecular interactions between nucleotides, a folding process that involves not only simpler secondary structures but also forces in three dimensions. In this process, genetic variation can directly affect the structure of coding RNA molecules [4,5], or non-coding RNA (ncRNA) involved in biological processes [6], such as microRNA (miRNA) [7], mitochondrial tRNAs (mt-tRNAs) [8,9], and long non-coding RNA (lncRNA) [10,11].

Protein is another common class of macromolecules, which play a crucial role in a series of biological processes such as cell proliferation, signal transduction, host-pathogen interaction, and protein transport. Minor changes in proteins can have dramatic effects on phenotypic results, such as the development of disease and drug resistance. Although the structure of protein is much easier to predict than RNA [12], Protein formation is more complex and diverse than RNA. Protein folding, stability, interaction, and activity may be affected by single-point and multi-point mutations. Among them, analyzing and determining the influence of genetic variation on amino acid sequence structure is a key step to explain the influence on protein structure.

Although little is known about how genetic variation ultimately leads to structural changes, the regulations can be found and summarized based on known molecular structure data and sequence data. In this process, the prediction of the macromolecular structure plays a significant role, and accurate prediction result can become an evaluation index of the structural impact caused by variation. Here, we aimed to review the influence of genetic variation on macromolecules, especially SNP and SNV on RNA and protein. some commonly used prediction methods and tools were summarized, as well as the impact of genetic variants on several clinical diseases.

Principle and process of RNA structure prediction

The function of RNA largely depends on its structure, which is affected by its sequence. However, the formation of structure is not a simple process, and RNA needs to be folded many times. In general, RNA sequences initially fold into the most thermodynamically favorable secondary structural elements, then forms a complementary paired double helix structure by self-folding [13]. Therefore, accurate prediction of secondary structure is a key step to predicting tertiary structure and function [14]. Some factors should be considered in predicting: First, there are about 40% non-canonical base pairs in nature which are base pairs other than A-U, G-C, and G-U [15]. Then, base triples are the cluster of three base interactions, which widely exist

in RNA structure [16] and can stabilize many RNA tertiary interactions [17]. Finally, pseudoknots will be formed when bases in different loops pair with each other [18]. These factors contribute to the functional diversity of RNA, but also increase the difficulty in predicting its structure.

Traditional RNA secondary structure prediction

In the era when direct measurement of RNA structure was not matured and RNA sequence data was scarce, and computational prediction was the mainstream method to identify RNA secondary structure. Most of these traditional methods were based on the minimum free energy (MFE) to simulate the RNA structure, but their accuracy and calculation speed met a bottleneck [19,20]. The optimal traditional method reached a performance ceiling of about 80% [21].

Traditional RNA secondary structure prediction methods can be divided into two categories: (i) Methods based on MFE score. This is also the most-used method, which is usually combined with the dynamic programming algorithm to find the MFE structure in the thermodynamic stable state. However, the calculating speed and accuracy on long sequence RNA and the prediction effect in the presence of pseudoknots are not ideal. (ii) A method based on comparative sequence analysis, which is more accurate than the former. Based on the assumption that RNA secondary structure is more conservative than RNA sequence in evolution, a group of homologous sequences is generally more similar in structure to predict secondary structure. Comparison sequence analysis can also predict structures with pseudoknots [22,23], but the accuracy is still limited. In addition, comparative sequence analysis can be combined with score-based methods [24-27]. However, a great limitation of this method is that many homologous sequences are in need.

Traditional RNA secondary structure prediction flows as follows Figure 1: First, RNA sequence data and its molecular thermodynamics and molecular dynamics parameters are used to obtain a series of possible free energy structures through a dynamic programming algorithm [28]. Secondly, the partition function and base pairing probabilities of RNA molecules are obtained according to the specific constraints of the actual environment such as temperature. Then, use them to trace back a series of possible structures obtained in the first step to get a new MFE structure. Finally, the optimal free energy structure is regarded as the result of prediction, and a graphical output in the form of an energy point diagram is formed. (Figure 1)

Machine learning-based RNA secondary structure prediction

In 2010, Kertesz et al. [20] described parallel analysis of RNA structure (PARS), which combined high-throughput technology with RNA structure-specific enzymes to provide secondary structure analysis of thousands of RNAs at single nucleotide resolution. Many detection methods have been developed and applied to RNA structure analysis, so the quality and quantity of RNA sequence data has been continuously improved. With the development and improvement of machine learning (ML) technology, especially deep learning, prediction methods based on ML have gradually developed and replaced traditional methods.

ML-based methods train their models in a supervised way [29]. These models learned to map features to structures by adjusting model parameters according to known structures and corresponding sequences and other feature information (Supplementary File S2.1). Many of them used free energy parameters, encoded RNA sequences, sequence patterns, or evolutionary information as key features, and their outputs can be classified as tags (such as paired or unpaired) or continuous values (such as free energy). When features of the new structure are input into the training model, the model can classify the corresponding tag or predict the corresponding parameter [29]. Its characteristic is that it contains all the information of the data, so it does not rely on the assumptions in traditional methods and is easy to combine with known biological rules [19]. While the accuracy of prediction is improved, the prediction model after training is faster than traditional methods, so it has more advantages in processing long RNA sequences.

The development history and representative methods of RNA prediction

Since the concept of RNA secondary structure was put forward in 1960 [30], the research on how RNA secondary structure formed from sequence to function has never stopped. In 1971, Tinoco et al. [31] proposed a simple method to estimate the secondary structure of RNA molecules for the first time, but Delisi et al. [32] didn't get the expected results when trying to predict the secondary structure of RNA from the MFE. In 1981, Zuker et al. [33] used a dynamic programming algorithm to predict RNA secondary structure based on the MFE model. In 1989, Zuker et al. developed a computer program, which calculates the structure of optimal and suboptimal folds and becomes the basis of the mfold package [34]. In 1994, Hofacker et al. [28] proposed the ViennaRNA package including the RNAfold method to calculate the MFE structure of RNA molecules or partition function [12]. In 2004, Ding et al. proposed the MFE model combined with statistical methods and developed the software package Sfold. In 2010, Halvorsen et al. developed the SNPfold algorithm which compares the sequence information before and after mutation based upon the calculation of the RNAfold partition function. [35]. In 2014, David H. Mathews provided four programs in the RNAstructure based on comparative sequence analysis. [36]. With the development of RNA structure detection technology, learning-based methods were increasingly used to predict RNA structure. In 2020, Chen et al. proposed an end-to-end deep learning model, called E2Efold, which significantly improved the accuracy and speed of prediction [37]. In 2022, Laiyi Fu et al. developed the UFold method based on deep learning, which further improved the accuracy of prediction [38]. The development history of the common RNA structure prediction method is shown in Figure 2.

Here we list several representative methods for RNA prediction and more details of the RNA traditional prediction software packages are shown in supplementary materials, including mfold package (Supplementary File S1.1), ViennaRNA package (Supplementary File S1.2), and RNAstructure package (Supplementary File S1.3).

Mfold was first proposed by Zuker et al. in the 1980s [34,39,40] , which introduced new improvements to make the calculation of RNA structure prediction more accurate and efficient. The first is the combination of the new energy rules, and the second is the ability to calculate optimal and suboptimal folding. Conditional constraints can be imposed based on preassumptions, and sub-optimal structures may be more consistent with experimental data than optimal ones. In terms of output, the best and suboptimal folding lists can be sorted by energy. An energy point diagram can also be used to describe all suboptimal folds in one image [41].

RNAfold was developed based on the ViennaRNA package, the computer program widely used to compute and compare RNA secondary structures. Using the dynamic programming algorithm, the code predicted the structure with MFE after calculating the equilibrium partition function and base pairing probabilities, which may serve as constraints. Based on this, RNAfold developed an RNA secondary structure algorithm based on tree editing and alignment, which can calculate the MFE, backtrack the optimal secondary structure and efficiently solved the RNA inverse folding problem. It was worth noting that RNAfold can provide constraints for the folding algorithm to force the pairing of a specific location [13,28].

SNPfold algorithm was designed based on RNA partition function calculation in RNAfold [26,42]. The difference between them is that the SNPfold algorithm requires two different RNA strands with the same length. One strand is a wild-type RNA sequence, and the other is an RNA sequence containing genetic variation. SNPfold will calculate the Pearson correlation coefficient between two RNA base pairs and the partition function of the possible RNA conformations set of the sequence to analyze the influence of SNPs on RNA structure. Besides, it will help to identify the disease-related mutation in the regulatory RNA by analyzing genome-wide association studies (GWAS) data and the whole mRNA structure [35].

UFold was a deep learning-based method recently developed to directly train labeled data and base pairing rules. A novel RNA sequence image representation method was proposed by UFold, which can be effectively processed by Fully Convolutional Networks (FCNs). It was found that it outperformed other methods in the family dataset and its prediction speed was improved [38].

Principle and process of protein structure prediction

Changes in protein structure are influenced by a variety of factors, including SNPs and mutations. Nucleotide variations that cause an amino acid change are called nonsynonymous mutations, and nonsynonymous SNP (nsSNPs) are an important part of this extensive research. These single-base changes, or multiple nucleotide substitutions leading to alterations in the amino acid sequence of the encoded protein, are called single amino acid variations (SAVs) or missense variations. Amino acid changes can affect protein stability, interactions, and enzyme activity, and even cause disease. Therefore, it is crucial to accurate prediction of the effect of genetic variation on protein structure for understanding the mechanism of genome variation associated with certain diseases [43].

By reviewing the literature, we found that the prediction of the effect of genetic variants on protein structure is always inseparable from machine learning methods. Prediction methods often combine protein characterization data, such as protein dynamics, contact potential scores, interatomic interactions, and other aspects features, using machine learning algorithms including support vector machines, random forests, and deep learning to train and implement the predictions. Therefore, several feature extraction methods and machine learning methods are reviewed in this section.

Protein structure feature extraction method

Normal Mode Analysis (NMA) provides a valuable method for the study of system dynamics and accessible conformations as an alternative to time-consuming and computationally expensive molecular dynamics simulation. The kinetic properties extracted from the protein structure generated from the NMA module of the Bio3d tool can be utilized [44].

Analysis of Mutation Effects. The change of folded Gibbs free energy may be caused by many related factors. To combine these characteristics, Arpeggio [45] can be used to calculate the number of hydrophobic contacts involving wild-type residues and the contact potential score in the AAINDEX database [46].

Graph-based structural signatures approach to represent molecular structures has proven to be successful for a range of applications towards the study of protein structure and changes carried out by missense mutations, including phenotypic changes. These signatures comprise physicochemical and geometrical properties from the wild-type environment that are based on distance patterns mined from the 3D structure by representing atoms as nodes and their interactions as edges. Then the physicochemical properties and the distance pattern between atoms are defined according to the properties of amino acids (i.e., pharmacophores) and transformed into a cumulative distribution function [47-50].

ML

ML was successfully applied to proteins earlier than RNA (Supplementary File S2.2). As executors of functions, proteins are robust and have a large number of features that can be used for ML. In the field of bioinformatics, algorithms such as deep learning, random forest, support vector machine, etc. are widely used in protein structure prediction [51], protein functional site prediction [52,53], subcellular location prediction, etc [54,55]. There are many ML algorithms, none of which is the best algorithm for all tasks. Take the method flow of the classic sequence-based tool DynaMut2 [56] as an example, as shown in the following figure 3.

The development history and representative methods of protein prediction

Several methods have been developed to predict how missense mutations affect protein stability by using sequence or structural information from which the information is often complementary. The development history of relevant methods is shown in Figure 4. According to the characteristics of existing methods, we divide them into five categories and introduce their respective characteristics and representative methods.

The structure-based method only (red part in the figure)

mCSM was developed by Pires et al. in 2014, which was a method based on protein structure and relies on the graph-based signature to study missense mutation [47]. It encodes the interatomic distance to represent the protein residue environment and trains the prediction model. Subsequent studies have also proven that the effect of mutation is related to the atomic distance around amino acid disability. The mCSM network server has been established and provides many extended tools and methods.

SDM is an algorithmic program based on statistical potential energy function proposed by Topham et al. in 1997 [57]. Worth et al. established the SDM network server in 2011 [58]. Based on the structure method, SDM uses the amino acid substitution frequency of the homologous protein family in different environments to calculate the stability score between wild-type and mutant proteins. The change in protein stability is one of the important parts to estimate the effect of genetic variation on protein structure.

The sequence-based method only (green part in the figure)

MuStab was a network server that was developed by Teng et al. in 2010 [59], which is a machine-learning method for detecting protein stability based on sequence characteristics. It uses experimental data on the free energy variation of protein stability during mutation. After analyzing 20 sequence features, it is found that the classifier combined with 6 sequence features is the most accurate including stability (S3) bulkiness (Bu), the transmembrane tendency (Tt), beta-sheet (B), average area buried on transfer from standard state to folded protein (Aa), and the mobility of an amino acid on chromatography paper (Mc) (Supplementary File S2.2).

Methods based on structure and sequence (yellow part in the figure)

Imutant3. 0 was proposed by Capriotti et al. in 2008 [60], which can distinguish the experimental protein stability free energy change value ($\Delta\Delta G$) into stable mutation, unstable mutation, and neutral mutation by using a support vector machine (SVM) based on sequence or structure. It also improves the prediction effect of free energy change caused by single-point protein mutation.

Other methods (purple part in the figure)

The **iStable** is an integrated predictor, which can predict the change of protein stability by SVM when a single amino acid residue is mutated. It uses sequence information and prediction results that adopted the SVM as an integrator from different element predictors to construct grid computing architecture [61]. The iStable2.0 systematically improves prediction performance based on iStable [62].

The new method (blue part in the figure)

Many new methods have emerged in recent years. Among them, AlphaFold2 [63] based on deep learning, and RoseTTAFold [64] using a 3-track neural network are two representative methods, both of which have achieved amazing accuracy in predicting protein structure.

The latest version of **Alphafold** is based on a new ML method, which integrates the physical and biological knowledge of protein structure into the design of deep learning algorithm by using multi-sequence alignment [63]. AlphaFold is a fully redesigned neural network-based model that has similar prediction accuracy to the experimental structure and significantly outperforms other methods in most cases.

RoseTTAFold was developed by Minkyungbaek and others [64]. Combined with the network architecture of relevant ideas, RoseTTAFold achieved the best performance by successively converting and integrating the information of the one-dimensional sequence layer, two-dimensional distance layer, and three-dimensional coordinate layer-3. The structure prediction accuracy of the track network is close to that of deepmind in the 14th round of the Critical Assessment of Structure Prediction (CASP14).

Common identification tools and software

Here, we list some tools that can predict the effect of mutation on the macromolecular structure (Supplementary Table T1). These tools combine the information related to mutation and macromolecular structure prediction methods. The tool usually takes the molecular sequence and variation information as input, the predicted wild-type structure and the structure after variation, and the change of macromolecular thermodynamic index as output. We have developed a website GEMS (*http://www.onethird-lab.com/gems/*) as a brief introduction and index to these tools.

Tools to identify structural effects on RNA

RNAsnp (*https://rth.dk/resources/rnasnp*) is a web server tool, which use a different mode to predict the effect of SNPs on different length of RNA sequence. The global folding method RNAfold [28] calculate the base pair probabilities of wild-type and mutant sequence, which is less 1000nt, and the local folding method RNAplfold [65] are used for large RNA sequence. These two methods are part of the ViennaRNA package [13], for more information see the annex. SNP effects are quantified from extensive pre-computed tables of distributions of substitution effects as a function of gene length and GC content. The input data of RNAsnp can be a single RNA sequence in FASTA format with one or more mutants whose structural effect needs to be predicted. It not only provides the structural prediction results but also features a graphical output representation [66,67] (Supplementary File S5.1).

MutaRNA(*http://rna.informatik.uni-freiburg.de/MutaRNA*) is a web server for studying SNV-induced RNA structure changes, which is also the first tool that provides different dot plots for comparative analysis of base pairing potentials. MutaRNA uses the local folding method RNAplfold [65] which is part of the ViennaRNA package [13] to retrieve candidate RNAs. MutaRNA also integrates the empirical p-values from RNAsnp [67] and the relative entropy comparing wild-type and mutant-form remuRNA [68] to quantify the structure aberration caused by SNV. The input data of MutaRNA is an RNA sequence of the wildtype sequence (WT) in FASTA format and the location of the mutation. It can provide a variety of visualization results, including heat map matrices, circular plots, and arc plots that are convenient for users to use directly in scientific reporting [69] (Supplementary File S5.2).

LncCASE (*http://bio-bigdata.hrbmu.edu.cn/LncCASE*) is a network database, which is constructed based on lncRNAs prediction in cancer. Multidimensional molecular analysis of tumor samples, biomolecular interaction networks, and pathway data resources by integrating genomic and transcriptome data from human cancer. LncCASE uses a computational method to identify the sub-pathways driven by lncRNAs under the influence of CNV. The copy number level of lncRNA was re-annotated, and the lncRNA CNV spectrum was constructed and visualized. The tool further analyzes the biological effects of lncRNA affected by genetic variation in cancer, which facilitates the study of cancer biology [70].

PON-mt-tRNA(*http://structure.bmc.lu.se/PON-mt-tRNA*) is a prediction tool for pathogenic variants on tRNAs. Since all pathogenic variants of tRNAs are located in mitochondria, investigators collected mt-tRNA variants and developed a machine-learning random forest algorithm based multivariate probabilistic prediction method. The method requires a reference position in the mtDNA, the reference (original) nucleotide, and the nucleotide altered by each variation as inputs. In addition, users have the option to submit evidence for isolation, biochemical, and histochemical characterization. The investigators classified all possible single nucleotide substitutions in all human mt-tRNA using PON-mt-tRNA, which documents the prediction of all possible nucleotide substitutions in mt-tRNA genes [71].

UFold (*https://ufold.ics.uci.edu*) is developed as a web server running UFold to facilitate the use of the UFold method. Users can enter or upload RNA sequences in FASTA format. The server predicts the secondary structure of RNA using pre-trained UFold models (trained on all datasets) and stores the predicted structure in a dot-bracket file or bpseq file for users to download. The user can also select in the options panel either to predict non-canonical pairs or not directly. The server further provides interface connectivity to the VARNA tool [72] to visualize the predicted structures. Most existing RNA prediction servers, such as

RNAfold [13], MXfold2 [73], and SPOT-RNA [74], can only predict one RNA sequence at a time and limit the length of the input sequence, but UFold does not have that limitation. Unfortunately, UFold is currently unable to combine SNP data to predict the effect of structure breaking mutations. More tools for mutation prediction of RNA structural disruption effects are shown in Supplementary File S3.

Tools of identifying structural effects on proteins

INPS (*http://inps.biocomp.unibo.it*) is a new approach that departs from protein sequence information and does not rely on structure to annotate the effect of non-synonymous mutations on protein stability. INPS is based on support vector machine regression and is trained to predict the change in thermodynamic free energy based on a single point change in a protein sequence. It has the advantage of being suitable for calculating the effect of nonsynonymous polymorphisms on protein stability when the protein structure is unavailable. INPS predictor consists of one support vector regression (SVR) trained on single point variations in different proteins and can complement each other [75] with methods like structure-based mCSM [47].

DynaMut2 (*http://biosig.unimelb.edu.au/dynamut2*) is a tool that integrates information on protein dynamics and structural environment attributes of wild-type residues. As a graph-based signature method, it is able to accurately predict the effect of mutations on the stability and dynamics of single and multiple point mutations. DynaMut2 can predict the Gibbs free energy change of single point mutations or no more than 3 multiple point mutations based on Normal Mode Analysis (NMA) and protein kinetic analysis. Its input may be a single mutation or a mutation list, and its performance is better than other methods in predicting stability changes caused by single-point mutations [56].

mCSM-PPI2 (*http://biosig.unimelb.edu.au/mcsm-ppi2/*) is a new machine-learning computational tool that can predict the effect of missense mutations on the binding affinity of protein interactions accurately. It leverages graph-based structural signatures to model inter-residual interaction networks, evolutionary information, complex network metrics, and the change effects of energy terms to generate optimized predictors. The mCSM-PPI2 can be used to assess the impact of user input of a specified mutation or to predict the impact of protein interface mutations in an automated manner [76].

PhyreRisk (*http://phyrerisk.bc.ic.ac.uk*) is a web application tool for connecting genomic, proteomic, and structural data to facilitate the mapping of human variants to protein structures. It provides information on 20,214 human typical protein sequences and 22,271 alternative protein sequences (isoforms) and supports new variant data in a genomic coordinate format (VCF, applying reference SNP IDs and HGVs release symbols) and human gene builds GRCh37 and GRCh38 as inputs. In addition, it supports the use of amino acid coordinates to map variations and search for genes or proteins of interest. PhyreRisk aims to enable researchers to translate genetic data into protein structural information that provides a more comprehensive assessment of the functional impact of variants [77].

AlphaFold (*https://alphafold.ebi.ac.uk*) is an open-access and extensive database that provides highly accurate protein structure prediction. AlphaFold V2.0 gives an unprecedented expansion of structural coverage of known protein sequence space structures. The latest version of AlphaFold is based on a novel ML approach that combines physical and biological knowledge about protein structure. Using multiple sequence alignments, this knowledge is incorporated into the design of deep learning algorithms. Not only that, AlphaFold2 also uses inductive biases in physics and geometry to build components that learn from PDB data. This enables the network to learn more efficiently from limited data in the PDB and to deal with the complexity and diversity of structural data. AlphaFold and its technology computational methods have been important tools to solve biophysical problems in modern biology [63]. However, it is a pity that up to now, no tool has been developed to predict the structural effects caused by mutation [78]. More protein structure disruption effect mutation prediction tools are shown in Supplementary File S4.

Table 1. The information on the protein prediction tools.

Tools	Methods&Algorithms	Input	Output

Tools	Methods&Algorithms	Input	Output
NAT22PRED	SVM predictor	SNP	Genotypes and
INPS	SVM regression	nsSNPs	Prediction from
DynaMut2	Graph-based signatures	Mutations	Protein dynami
mCSM-PPI2	Graph-based signatures	Single mutation	Structure predi
PhyreRisk	Homology modeling	Genomic or protein variants	Structure predi
AlphaFold2	Deep-learning	Protein name, gene name, UniProtaccession, or organism name	Structure predi

Other identification tools

IntSplice(*http://www.med.nagoya-u.ac.jp/neurogenetics/IntSplice*) is the web server to identify SNVs affecting intronic cis-elements. The precise spatiotemporal regulation of splicing is mediated by the splicing cis-elements on pre-mRNA. IntSplice uses an online SVM model based on the analysis of the effect size of each intron nucleotide on the annotated alternative splicing. It can predict the splicing consequences of SNVs at intron positions -50 to -3 in the human genome. The IntSplice model was applied to distinguish pathogenic SNVs from the Human Gene Mutation Database and normal SNVs from the dbSNP database and achieved good results [79].

Natural Language Processing based non-synonymous Single Nucleotide Polymorphism Predictor (**NLP-SNPPred,**http://www.nlp-snppred.cbrlab.org.) web server could distinguish pathogenic protein-coding variations and neutral protein-coding variations based on the state-of-the-art Natural language Processing (NLP) of Artificial Intelligence (AI). Through feature extraction, a multi-class classifier (CLF1) followed by a binary-class classifier (CLF2) is created. NLP-SNPPred uses the NPL approach to read biological literature for the identification of pathogenic versus neutral variants and outperforms state-of-the-art functional prediction methods and can be used to predict functional effects of protein-coding mutations. NLP will add more features such as homology, epigenomics, and evolutionary information [80].

The impact of genetic variation on macromolecular structure in diseases

Although GWAS has identified some genetic variants, the mechanisms between genetic variation and disease remain unclear. Genetic variations can alter the structure of RNA or proteins, leading to abnormalities in biological function or even disease. Some studies found that the molecular mechanism of some typical diseases was closely related to the structural effects of genetic variation, such as hypertension [81,82], Retinoblastoma [83,84], β Thalassemia [85-90], and so on.

Hereditary hyperferritinemia-cataract syndrome

Hereditary hyperferritinemia-cataract syndrome (HHCS) is a rare disease characterized by high serum ferritin levels and congenital bilateral cataracts. U22G and U22G - G14C are two SNPs in the 5'- UTR of ferritin light (FTL) mRNA. FTL chain is an iron-responsive element (IRE) in 5'- UTR, which plays a major regulatory role in mRNA translation. Some studies found that [91] these two SNPs can affect RNA structure and subsequent gene function. SNP U22G can disrupt the structure of the IRE, leading to abnormal FTL gene regulation. However, U22G - G14C can restore the mutated ire to wild type [92]. Rs886037623 (T22G) changes the spatial structure of mRNA folding because the original U was replaced by G in mRNA [93]. Under normal circumstances, in a low iron environment, iron regulatory proteins (IRP) will combine with IRE in correctly folded mRNA to form a repressor complex of protein synthesis, and the synthesis of ferritin is inhibited. After mutation, the structurally altered mRNA can no longer bind to IRP, the transcriptional regulation is lost, and a large amount of ferritin is secreted, resulting in the formation of hyperferritinemia.

At the same time, too much ferritin precipitates in the lens, resulting in cataracts [94]. The effects of genetic variation on HHCS are shown in Figure 5A.

Sicklemia

Sicklemia is the most serious of the abnormal hemoglobinopathy, whose clinical manifestations are chronic hemolytic anemia, susceptibility to infection, and chronic ischemia leading to organ and tissue damage [95]. Its pathogenesis is complex and significantly related to genetic factors. There is evidence that the T is replaced by A, in rs334 on the gene encoding hemoglobin, after the transcription and translation process, glutamic acid is replaced by valine to form abnormal hemoglobin at the sixth position in the β-chain amino acid sequence [96]. When the oxygen partial pressure decreases, hemoglobin molecules interact with each other to form a spiral polymer, which distorts red blood cells into sickle cells, and finally leads to anemia [97]. The effects of genetic variation on Sicklemia are shown in Figure 5B.

The effects of genetic variation on HHCS are shown in Figure 5A. Because of the mutation rs886037623, T -> G, and the corresponding U in the transcribed mRNA is replaced by G, leading to its structural change, IRP cannot bind to it, and finally resulting in the overexpression of ferritin. The effects of genetic variation on Sicklemia are shown in Figure 5B. Because of the mutation rs334, T > A, and through the subsequent transcription and translation process, Glu (E) at position 6 of the amino acid sequence of the protein becomes Val (V). Change the structure of hemoglobin molecules, and finally lead to Sicklemia.

Tumor and cancer

In cancer research, the influence of genetic variation on the macromolecular structure cannot be ignored. For example, Retinoblastoma (RB) is a malignant tumor caused by photoreceptor precursor cells, which is common in children under 3 years old and has family genetic susceptibility [98]. It is proven that some mutations are closely related to RB [84]. J K Cowell et al. [83] identified a novel mutation(G-C) within a core motif of specificity protein 1 (SP1) transcription factor from a family with a mild RB and a band shift of an unidentified protein was found in the mutant oligomer. This protein may affect the expression of the RB1 gene and eventually lead to RB.

In addition, the influence of genetic variation on lncRNA has been extensively explored in some cancers. LINC00673 is a potential tumor suppressor of pancreatic cancer. Rs11655237 is an SNP in the exon of LINC00673, which causes LINC00673 to have a new binding target, thus weakening its role and increasing the risk of pancreatic cancer [99]. A713G and T714C mutations in lncRNA GAS8-AS1 accelerate the growth of cancer cells and increase the risk of thyroid cancer [100]. Abnormal copy number and expression of somatic cells on focally amplified lncRNA on chromosome 1 (FAL1) can inhibit *P21*and lead to ovarian cancer [11].

COVID-19

Since 2019, COVID-19 has become/been a major threat to global health. Patients infected with SARS-CoV-2 may have not only acute respiratory distress syndrome [101], but also acute heart injury, heart failure, inflammation leading to sepsis, and multiple organ dysfunction [102]. As an RNA virus, its replication process is prone to mutation. Although most structural proteins of SARS-CoV-2 are conserved in the corona virus family, the sequence similarity can reach 90%. However, small sequence changes will have a huge impact on the structure and pathogenesis of SARS-CoV-2. For example, the N501Y substitution of spike protein increases the binding affinity between RBD and ACE2 receptor in the structural proteins that make up SARS-CoV-2, thus improving the transmission rate [103,104]. The P71L mutation of envelope protein was statistically associated with disease severity and mortality [105].

Other diseases

Parkinson's disease (PD) is a common neurodegenerative disease in the elderly, which is caused by both genetic and environmental factors. DJ-1 is a protein-coding gene whose loss of function can lead to neurodegeneration. Some studies have proven that DJ-1 mutation is associated with the susceptibility gene *PARK7* in PD [106]. L166P is the most typical DJ-1 missense mutation, which affects the integrity of -helix G,

resulting in poor protein folding in PD [107,108]. It has been suggested that genetic variation in the DJ-1 protein can affect the structure of the protein and therefore the occurrence and development of PD.

Amyotrophic lateral sclerosis (ALS), also known as motor neuron disease (MND), is the progressive muscle weakness and atrophy caused by the injury of upper motor neurons and lower motor neurons [109]. Some studies have shown that the mutation of human superoxide dismutase (HSOD) is related to the familial heritability of ALS [110]. The rs121912442 transforms Ala into Val in the fourth position of the amino acid sequence, which leads to the side chain shift of the protein, destroys the stability of the dimer interface, reduces the enzyme activity of the mutant protein compared with the wild type, and promotes the formation of HSOD-containing aggregates that are toxic to motor neurons [111]. This suggested that the change of macromolecular structure caused by mutation could change the enzyme activity and lead to the aggregation of mutant proteins and even toxic proteins [110,112].

Congenital myasthenia syndrome (CMS) is a heterogeneous neuromuscular disease characterized by muscle weakness. Its pathogenesis is complex and is caused by genetic factors and environmental factors [113]. It is generally accepted that genetic factors, especially mutations in some important genes, play a key role in CMS [114]. Akihide et al. found that the homozygous mutation (C.913-5T > A) in RAPSN could inhibit the binding of U2AF65 and splicing cis-elements and disrupt the splicing process, leading to CMS [79]. Saito et al. identified a missense mutation c.737C > T (p.A246V) in RAPSN as an important factor in CMS by exome sequencing [113].

Charcot-Marie-Tooth (CMT) is one of the most common hereditary peripheral neuropathies (incidence rate is about 1/2500), which is characterized by progressive muscle weakness and atrophy of the distal extremities with sensory disorders [115]. Studies have found that the mutation rs137852646 caused the amino acid variation G526R of human glycyl-tRNA synthetase (GlyRS), and the dimer formed by G526R protein is tighter than that of the wild type, blocking the binding of GlyRS amp active site, resulting in the loss of aminoacylation activity [106]. Therefore, the change of dimer interface caused by genetic variation may be an important factor of CMT, which helps us understand the CMT mechanism [116].

Discussion

Genetic variation has a complex effect on the structure of macromolecules, and it is of great significance to predict this effect. We know that structure determines the function of macromolecules in biological processes, and effective and accurate predictions would greatly accelerate the research process of disease mechanisms under the influence of genetic variation on macromolecules with abnormal structures. The exploration of structures has gradually introduced dynamic programming algorithms, partition functions, sequence alignment, and ML from the initial MFE model based only on thermodynamic laws, which has continuously broadened the dimension of describing structural changes, and finally achieved the current more accurate and efficient results. The tools combined with the corresponding methods and genetic variation have been used to predict the structural effects due to mutations, and the results can explain some disease mechanisms.

At present, many methods and tools are commonly used to predict the impact of SNP on macromolecules, each with advantages and disadvantages. The traditional RNA prediction methods are relatively mature, but the accuracy and speed of prediction are not as good as the new methods combined with ML. The influence of genetic variation on macromolecular structure will be more explained under the new method. Stunning progress has been made in protein structure prediction. AlphaFold2 developed by deepmind has achieved high accuracy in predicting protein structure [63]. Minkyung Baek et al. obtained the best performance by combining with the three-track network architecture, the accuracy of structure prediction was close to that of deepmind in CASP14 [64]. However, the combination of genetic variation information and these methods to predict the impact of genetic variation on protein structure is still in the initial stage [78]. As the accuracy of calculations increases, the details of the prediction of biological macromolecules are becoming clearer. People are trying to quantify the impact of genetic variation on macromolecular structure [112,117].

At the same time, the prediction objects also range from RNA with shorter sequence structures to longer and more complex chromosome structures [118].

The prediction of the influence of genetic variation on macromolecular structure will become increasingly accurate as computational efficiencies increase and prediction methods diversify. It is hoped that the application of the ML method to predict the structural effects caused by genetic variation will bring new understanding to complex diseases and cancers. It is believed that with the increase of the accuracy and breadth of structure prediction, people will be able to establish a set of personalized birth time sequence data in the future, predict the disease risk caused by the changes of some macromolecular structures of individuals according to the genetic variation, and give corresponding suggestions. In addition, differences in genetic variation among populations have potential effects on macromolecular structures, such as the efficacy of drugs on specific targets varies from person to person. Changes in the corresponding target can be predicted by studying genetic variation in populations or even individual patients. It is even possible to formulate a reasonable individualized medication plan by predicting the structural effect of the individual mutation, to achieve the goal of precision medical treatment.

Acknowledgements

This work was supported by the National Natural Science Foundation of China [Grant Nos. 31970651, 92046018]; Marshal Initiative Funding NO. HUMMIF-22010. Mathematical Tianyuan Fund of the National Natural Science Foundation of China [Grant No. 12026414].

Competing interests

The authors declare no competing interests.

Author Contributions

J.X.K., S.Y.W., Z.J., Y.N.M., and H.Y.C contributed equally to this work. Y.S.J., H.C.L, M.M.Z. conceived and contributed the work. C.S., J.X, J.X.T., Y.D., W.H.L., H.S.T., X.Y.G., drafted and modified the manuscript. S.B., C.Z. are important contributors of the GWAS Project. The GWAS Project provided data support.

References

1 Ennis, S., Murray, A., Brightwell, G., Morton, N. E., & Jacobs, P. A. (2007). Closely linked cis-acting modifier of expansion of the CGG repeat in high risk FMR1 haplotypes. *Hum Mutat, 28* (12), 1216-1224. doi:10.1002/humu.20600

2 Pearson, C. E., Nichol Edamura, K., & Cleary, J. D. (2005). Repeat instability: mechanisms of dynamic mutations. *Nat Rev Genet, 6* (10), 729-742. doi:10.1038/nrg1689

3 Gerhardt, J., Zaninovic, N., Zhan, Q., Madireddy, A., Nolin, S. L., Ersalesi, N., et al. (2014). Cis-acting DNA sequence at a replication origin promotes repeat expansion to fragile X full mutation. *J Cell Biol, 206* (5), 599-607. doi:10.1083/jcb.201404157

4 Chen, J. M., Ferec, C., & Cooper, D. N. (2006). A systematic analysis of disease-associated variants in the 3' regulatory regions of human protein-coding genes II: the importance of mRNA secondary structure in assessing the functionality of 3' UTR variants. *Hum Genet, 120* (3), 301-333. doi:10.1007/s00439-006-0218-x

5 Shen, L. X., Basilion, J. P., & Stanton, V. P., Jr. (1999). Single-nucleotide polymorphisms can cause different structural folds of mRNA. *Proc Natl Acad Sci U S A, 96* (14), 7871-7876. doi:10.1073/pnas.96.14.7871

6 Ramirez-Bello, J., & Jimenez-Morales, M. (2017). [Functional implications of single nucleotide polymorphisms (SNPs) in protein-coding and non-coding RNA genes in multifactorial diseases]. *Gac Med Mex, 153* (2), 238-250.

7 Haas, U., Sczakiel, G., & Laufer, S. D. (2012). MicroRNA-mediated regulation of gene expression is affected by disease-associated SNPs within the 3'-UTR via altered RNA structure. *RNA Biol, 9* (6), 924-937. doi:10.4161/rna.20497

8 Wittenhagen, L. M., & Kelley, S. O. (2003). Impact of disease-related mitochondrial mutations on tRNA structure and function. *Trends Biochem Sci, 28* (11), 605-611. doi:10.1016/j.tibs.2003.09.006

9 Tanaka, M., Takeyasu, T., Fuku, N., Li-Jun, G., & Kurata, M. (2004). Mitochondrial genome single nucleotide polymorphisms and their phenotypes in the Japanese. *Ann N Y Acad Sci, 1011* , 7-20. doi:10.1007/978-3-662-41088-2_2

10 Zhang, F., & Lupski, J. R. (2015). Non-coding genetic variants in human disease. *Hum Mol Genet, 24* (R1), R102-110. doi:10.1093/hmg/ddv259

11 Hu, X., Feng, Y., Zhang, D., Zhao, S. D., Hu, Z., Greshock, J., et al. (2014). A functional genomic approach identifies FAL1 as an oncogenic long noncoding RNA that associates with BMI1 and represses p21 expression in cancer. *Cancer Cell, 26* (3), 344-357. doi:10.1016/j.ccr.2014.07.009

12 McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure.*Biopolymers, 29* (6-7), 1105-1119. doi:10.1002/bip.360290621

13 Lorenz, R., Bernhart, S. H., Honer Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., et al. (2011). ViennaRNA Package 2.0.*Algorithms Mol Biol, 6* , 26. doi:10.1186/1748-7188-6-26

14 Zuker, M. (1989). Computer prediction of RNA structure. *Methods Enzymol, 180* , 262-288. doi:10.1016/0076-6879(89)80106-5

15 Leontis, N. B., & Westhof, E. (2001). Geometric nomenclature and classification of RNA base pairs. *RNA, 7* (4), 499-512. doi:10.1017/s1355838201002515

16 Abu Almakarem, A. S., Petrov, A. I., Stombaugh, J., Zirbel, C. L., & Leontis, N. B. (2012). Comprehensive survey and geometric classification of base triples in RNA structures. *Nucleic Acids Res, 40* (4), 1407-1423. doi:10.1093/nar/gkr810

17 Doherty, E. A., Batey, R. T., Masquida, B., & Doudna, J. A. (2001). A universal mode of helix packing in RNA. *Nat Struct Biol, 8* (4), 339-343. doi:10.1038/86221

18 van Batenburg, F. H., Gultyaev, A. P., & Pleij, C. W. (2001). PseudoBase: structural information on RNA pseudoknots. *Nucleic Acids Res, 29* (1), 194-195. doi:10.1093/nar/29.1.194

19 Zhao, Q., Zhao, Z., Fan, X., Yuan, Z., Mao, Q., & Yao, Y. (2021). Review of machine learning methods for RNA secondary structure prediction. *PLoS Comput Biol, 17* (8), e1009291. doi:10.1371/journal.pcbi.1009291

20 Xu, B., Zhu, Y., Cao, C., Chen, H., Jin, Q., Li, G., et al. (2022). Recent advances in RNA structurome. *Sci China Life Sci, 65* (7), 1285-1324. doi:10.1007/s11427-021-2116-2

21 Seetin, M. G., & Mathews, D. H. (2012). RNA structure prediction: an overview of methods. *Methods Mol Biol, 905* , 99-122. doi:10.1007/978-1-61779-949-5_8

22 Engelen, S., & Tahi, F. (2010). Tfold: efficient in silico prediction of non-coding RNA secondary structures. *Nucleic Acids Res, 38* (7), 2453-2466. doi:10.1093/nar/gkp1067

23 Bellaousov, S., & Mathews, D. H. (2010). ProbKnot: fast prediction of RNA secondary structure including pseudoknots. *RNA, 16* (10), 1870-1880. doi:10.1261/rna.2125310

24 Ruan, J., Stormo, G. D., & Zhang, W. (2004). An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics, 20* (1), 58-66. doi:10.1093/bioinformatics/btg373

25 Hofacker, I. L., Fekete, M., Flamm, C., Huynen, M. A., Rauscher, S., Stolorz, P. E., et al. (1998). Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucleic Acids Res, 26* (16), 3825-3836. doi:10.1093/nar/26.16.3825

26 Bindewald, E., & Shapiro, B. A. (2006). RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. *RNA, 12* (3), 342-352. doi:10.1261/rna.2164906

27 Legendre, A., Angel, E., & Tahi, F. (2018). Bi-objective integer programming for RNA secondary structure prediction with pseudoknots.*BMC Bioinformatics, 19* (1), 13. doi:10.1186/s12859-018-2007-7

28 Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., & Schuster, P. (1989). Fast folding and comparison of RNA secondary structures %J Monatshefte für Chemie / Chemical Monthly.*125* (2).

29 Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349* (6245), 255-260. doi:10.1126/science.aaa8415

30 Fresco, J. R., Alberts, B. M., & Doty, P. (1960). Some molecular details of the secondary structure of ribonucleic acid. *Nature, 188* , 98-101. doi:10.1038/188098a0

31 Tinoco, I., Jr., Uhlenbeck, O. C., & Levine, M. D. (1971). Estimation of secondary structure in ribonucleic acids. *Nature, 230* (5293), 362-367. doi:10.1038/230362a0

32 Delisi, C., & Crothers, D. M. (1971). Prediction of RNA secondary structure. *Proc Natl Acad Sci U S A, 68* (11), 2682-2685. doi:10.1073/pnas.68.11.2682

33 Zuker, M., & Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information.*Nucleic Acids Res, 9* (1), 133-148. doi:10.1093/nar/9.1.133

34 Zuker, M. (1989). On finding all suboptimal foldings of an RNA molecule. *Science, 244* (4900), 48-52. doi:10.1126/science.2468181

35 Halvorsen, M., Martin, J. S., Broadaway, S., & Laederach, A. (2010). Disease-associated mutations that alter the RNA structural ensemble.*PLoS Genet, 6* (8), e1001074. doi:10.1371/journal.pgen.1001074

36 Mathews, D. H. (2014). Using the RNAstructure Software Package to Predict Conserved RNA Structures. *Curr Protoc Bioinformatics, 46* , 12 14 11-22. doi:10.1002/0471250953.bi1204s46

37 Chen, X., Li, Y., Umarov, R., Gao, X., & Song, L. J. a. e.-p. (2020). RNA Secondary Structure Prediction By Learning Unrolled Algorithms. arXiv:2002.05810.*https://ui.adsabs.harvard.edu/abs/2020arXiv200205810C*

38 Fu, L., Cao, Y., Wu, J., Peng, Q., Nie, Q., & Xie, X. (2022). UFold: fast and accurate RNA secondary structure prediction with deep learning.*Nucleic Acids Res, 50* (3), e14. doi:10.1093/nar/gkab1074

39 Jaeger, J. A., Turner, D. H., & Zuker, M. (1989). Improved predictions of secondary structures for RNA. *Proc Natl Acad Sci U S A, 86* (20), 7706-7710. doi:10.1073/pnas.86.20.7706

40 Jaeger, J. A., Turner, D. H., & Zuker, M. (1990). Predicting optimal and suboptimal secondary structure for RNA. *Methods Enzymol, 183* , 281-306. doi:10.1016/0076-6879(90)83019-6

41 Zuker, M. (1994). Prediction of RNA secondary structure by energy minimization. *Methods Mol Biol, 25* , 267-294. doi:10.1385/0-89603-276-0:267

42 Hofacker, I. L., & Stadler, P. F. (2006). Memory efficient folding algorithms for circular RNA secondary structures. *Bioinformatics, 22* (10), 1172-1176. doi:10.1093/bioinformatics/btl023

43 Lahti, J. L., Tang, G. W., Capriotti, E., Liu, T., & Altman, R. B. (2012). Bioinformatics and variability in drug response: a protein structural perspective. *J R Soc Interface, 9* (72), 1409-1437. doi:10.1098/rsif.2011.0843

44 Grant, B. J., Rodrigues, A. P., ElSawy, K. M., McCammon, J. A., & Caves, L. S. (2006). Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics, 22* (21), 2695-2696. doi:10.1093/bioinformatics/btl461

45 Jubb, H. C., Higueruelo, A. P., Ochoa-Montano, B., Pitt, W. R., Ascher, D. B., & Blundell, T. L. (2017). Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *J Mol Biol, 429* (3), 365-371. doi:10.1016/j.jmb.2016.12.004

46 Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., & Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res, 36* (Database issue), D202-205. doi:10.1093/nar/gkm998

47 Pires, D. E., Ascher, D. B., & Blundell, T. L. (2014). mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics, 30* (3), 335-342. doi:10.1093/bioinformatics/btt691

48 Pires, D. E., Blundell, T. L., & Ascher, D. B. (2015). pkCSM: Predicting Small-Molecule Pharmacokinetic and Toxicity Properties Using Graph-Based Signatures. *J Med Chem, 58* (9), 4066-4072. doi:10.1021/acs.jmedchem.5b00104

49 Pires, D. E. V., & Ascher, D. B. (2020). mycoCSM: Using Graph-Based Signatures to Identify Safe Potent Hits against Mycobacteria. *J Chem Inf Model, 60* (7), 3450-3456. doi:10.1021/acs.jcim.0c00362

50 Kaminskas, L. M., Pires, D. E. V., & Ascher, D. B. (2019). dendPoint: a web resource for dendrimer pharmacokinetics investigation and prediction. *Sci Rep, 9* (1), 15465. doi:10.1038/s41598-019-51789-3

51 Kedarisetti, K. D., Kurgan, L., & Dick, S. (2006). Classifier ensembles for protein structural class prediction with varying homology.*Biochem Biophys Res Commun, 348* (3), 981-988. doi:10.1016/j.bbrc.2006.07.141

52 Kumari, B., Kumar, R., & Kumar, M. (2014). PalmPred: an SVM based palmitoylation prediction method using sequence profile information.*PLoS One, 9* (2), e89246. doi:10.1371/journal.pone.0089246

53 Xu, Y., Ding, J., & Wu, L. Y. (2016). iSulf-Cys: Prediction of S-sulfenylation Sites in Proteins with Physicochemical Properties of Amino Acids. *PLoS One, 11* (4), e0154237. doi:10.1371/journal.pone.0154237

54 Liang, Y., Liu, S., & Zhang, S. (2016). Detrended cross-correlation coefficient: Application to predict apoptosis protein subcellular localization. *Math Biosci, 282* , 61-67. doi:10.1016/j.mbs.2016.09.019

55 Liu, T., Tao, P., Li, X., Qin, Y., & Wang, C. (2015). Prediction of subcellular location of apoptosis proteins combining tri-gram encoding based on PSSM and recursive feature elimination. *J Theor Biol, 366* , 8-12. doi:10.1016/j.jtbi.2014.11.010

56 Rodrigues, C. H. M., Pires, D. E. V., & Ascher, D. B. (2021). DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. *Protein Sci, 30* (1), 60-69. doi:10.1002/pro.3942

57 Topham, C. M., Srinivasan, N., & Blundell, T. L. (1997). Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables.*Protein Eng, 10* (1), 7-21. doi:10.1093/protein/10.1.7

58 Worth, C. L., Preissner, R., & Blundell, T. L. (2011). SDM–a server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Res, 39* (Web Server issue), W215-222. doi:10.1093/nar/gkr363

59 Teng, S., Srivastava, A. K., & Wang, L. (2010). Sequence feature-based prediction of protein stability changes upon amino acid substitutions. *BMC Genomics, 11 Suppl 2* , S5. doi:10.1186/1471-2164-11-S2-S5

60 Capriotti, E., Fariselli, P., Rossi, I., & Casadio, R. (2008). A three-state prediction of single point mutations on protein stability changes. *BMC Bioinformatics, 9 Suppl 2* , S6. doi:10.1186/1471-2105-9-S2-S6

61 Chen, C. W., Lin, J., & Chu, Y. W. (2013). iStable: off-the-shelf predictor integration for predicting protein stability changes.*BMC Bioinformatics, 14 Suppl 2* , S5. doi:10.1186/1471-2105-14-S2-S5

62 Chen, C. W., Lin, M. H., Liao, C. C., Chang, H. P., & Chu, Y. W. (2020). iStable 2.0: Predicting protein thermal stability changes by integrating various characteristic modules. *Comput Struct Biotechnol J, 18* , 622-630. doi:10.1016/j.csbj.2020.02.021

63 Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596* (7873), 583-589. doi:10.1038/s41586-021-03819-2

64 Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science, 373* (6557), 871-876. doi:10.1126/science.abj8754

65 Bernhart, S. H., Hofacker, I. L., & Stadler, P. F. (2006). Local RNA base pairing probabilities in large sequences. *Bioinformatics, 22* (5), 614-615. doi:10.1093/bioinformatics/btk014

66 Sabarinathan, R., Tafer, H., Seemann, S. E., Hofacker, I. L., Stadler, P. F., & Gorodkin, J. (2013). The RNAsnp web server: predicting SNP effects on local RNA secondary structure. *Nucleic Acids Res, 41* (Web Server issue), W475-479. doi:10.1093/nar/gkt291

67 Sabarinathan, R., Tafer, H., Seemann, S. E., Hofacker, I. L., Stadler, P. F., & Gorodkin, J. (2013). RNAsnp: efficient detection of local RNA secondary structure changes induced by SNPs. *Hum Mutat, 34* (4), 546-556. doi:10.1002/humu.22273

68 Salari, R., Kimchi-Sarfaty, C., Gottesman, M. M., & Przytycka, T. M. (2013). Sensitive measurement of single-nucleotide polymorphism-induced changes of RNA conformation: application to disease studies.*Nucleic Acids Res, 41* (1), 44-53. doi:10.1093/nar/gks1009

69 Miladi, M., Raden, M., Diederichs, S., & Backofen, R. (2020). MutaRNA: analysis and visualization of mutation-induced changes in RNA structure. *Nucleic Acids Res, 48* (W1), W287-W291. doi:10.1093/nar/gkaa331

70 Xu, Y., Wu, T., Li, F., Dong, Q., Wang, J., Shang, D., et al. (2020). Identification and comprehensive characterization of lncRNAs with copy number variations and their driving transcriptional perturbed subpathways reveal functional significance for cancer. *Brief Bioinform, 21* (6), 2153-2166. doi:10.1093/bib/bbz113

71 Niroula, A., & Vihinen, M. (2016). PON-mt-tRNA: a multifactorial probability-based method for classification of mitochondrial tRNA variations. *Nucleic Acids Res, 44* (5), 2020-2027. doi:10.1093/nar/gkw046

72 Darty, K., Denise, A., & Ponty, Y. (2009). VARNA: Interactive drawing and editing of the RNA secondary structure.*Bioinformatics, 25* (15), 1974-1975. doi:10.1093/bioinformatics/btp250

73 Sato, K., Akiyama, M., & Sakakibara, Y. (2021). RNA secondary structure prediction using deep learning with thermodynamic integration.*Nat Commun, 12* (1), 941. doi:10.1038/s41467-021-21194-4

74 Singh, J., Hanson, J., Paliwal, K., & Zhou, Y. (2019). RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nat Commun, 10* (1), 5407. doi:10.1038/s41467-019-13395-9

75 Fariselli, P., Martelli, P. L., Savojardo, C., & Casadio, R. (2015). INPS: predicting the impact of non-synonymous variations on protein stability from sequence. *Bioinformatics, 31* (17), 2816-2821. doi:10.1093/bioinformatics/btv291

76 Rodrigues, C. H. M., Myung, Y., Pires, D. E. V., & Ascher, D. B. (2019). mCSM-PPI2: predicting the effects of mutations on protein-protein interactions. *Nucleic Acids Res, 47* (W1), W338-W344. doi:10.1093/nar/gkz383

77 Ofoegbu, T. C., David, A., Kelley, L. A., Mezulis, S., Islam, S. A., Mersmann, S. F., et al. (2019). PhyreRisk: A Dynamic Web Application to Bridge Genomics, Proteomics and 3D Structural Data to Guide Interpretation of Human Genetic Variants. *J Mol Biol, 431* (13), 2460-2466. doi:10.1016/j.jmb.2019.04.043

78 Buel, G. R., & Walters, K. J. (2022). Can AlphaFold2 predict the impact of missense mutations on structure? *Nat Struct Mol Biol, 29* (1), 1-2. doi:10.1038/s41594-021-00714-2

79 Shibata, A., Okuno, T., Rahman, M. A., Azuma, Y., Takeda, J., Masuda, A., et al. (2016). IntSplice: prediction of the splicing consequences of intronic single-nucleotide variations in the human genome. *J Hum Genet, 61* (7), 633-640. doi:10.1038/jhg.2016.23

80 Rehmat, N., Farooq, H., Kumar, S., Ul Hussain, S., & Naveed, H. (2020). Predicting the pathogenicity of protein coding mutations using Natural Language Processing. *Annu Int Conf IEEE Eng Med Biol Soc, 2020* , 5842-5846. doi:10.1109/EMBC44109.2020.9175781

81 Inoue, I., Nakajima, T., Williams, C. S., Quackenbush, J., Puryear, R., Powers, M., et al. (1997). A nucleotide substitution in the promoter of human angiotensinogen is associated with essential hypertension and affects basal transcription in vitro. *J Clin Invest, 99* (7), 1786-1797. doi:10.1172/JCI119343

82 Ishigami, T., Umemura, S., Tamura, K., Hibi, K., Nyui, N., Kihara, M., et al. (1997). Essential hypertension and 5' upstream core promoter region of human angiotensinogen gene. *Hypertension, 30* (6), 1325-1330. doi:10.1161/01.hyp.30.6.1325

83 Cowell, J. K., Bia, B., & Akoulitchev, A. (1996). A novel mutation in the promotor region in a family with a mild form of retinoblastoma indicates the location of a new regulatory domain for the RB1 gene. *Oncogene, 12* (2), 431-436.

84 Macias, M., Dean, M., Atkinson, A., Jimenez-Morales, S., Garcia-Vazquez, F. J., Saldana-Alvarez, Y., et al. (2008). Spectrum of RB1 gene mutations and loss of heterozygosity in Mexican patients with retinoblastoma: identification of six novel mutations. *Cancer Biomark, 4* (2), 93-99. doi:10.3233/cbm-2008-4205

85 Jankovic, L., Efremov, G. D., Petkov, G., Kattamis, C., George, E., Yang, K. G., et al. (1990). Two novel polyadenylation mutations leading to beta(+)-thalassemia. *Br J Haematol, 75* (1), 122-126. doi:10.1111/j.1365-2141.1990.tb02627.x

86 Ho, P. J., Rochette, J., Fisher, C. A., Wonke, B., Jarvis, M. K., Yardumian, A., et al. (1996). Moderate reduction of beta-globin gene transcript by a novel mutation in the 5' untranslated region: a study of its interaction with other genotypes in two families. *Blood, 87* (3), 1170-1178.

87 Ho, P. J., Hall, G. W., Luo, L. Y., Weatherall, D. J., & Thein, S. L. (1998). Beta-thalassaemia intermedia: is it possible consistently to predict phenotype from genotype? *Br J Haematol, 100* (1), 70-78. doi:10.1046/j.1365-2141.1998.00519.x

88 Kazazian, H. H., Jr., & Boehm, C. D. (1988). Molecular basis and prenatal diagnosis of beta-thalassemia. *Blood, 72* (4), 1107-1116.

89 Waye, J. S., Eng, B., Patterson, M., Reis, M. D., Macdonald, D., & Chui, D. H. (2001). Novel beta-thalassemia mutation in a beta-thalassemia intermedia patient. *Hemoglobin, 25* (1), 103-105. doi:10.1081/hem-100103075

90 Morgado, A., Picanco, I., Gomes, S., Miranda, A., Coucelo, M., Seuanes, F., et al. (2007). Mutational spectrum of delta-globin gene in the Portuguese population. *Eur J Haematol, 79* (5), 422-428. doi:10.1111/j.1600-0609.2007.00949.x

91 Hentze, M. W., & Kuhn, L. C. (1996). Molecular control of vertebrate iron metabolism: mRNA-based regulatory circuits operated by iron, nitric oxide, and oxidative stress. *Proc Natl Acad Sci U S A, 93* (16), 8175-8182. doi:10.1073/pnas.93.16.8175

92 Martin, J. S., Halvorsen, M., Davis-Neulander, L., Ritz, J., Gopinath, C., Beauregard, A., et al. (2012). Structural effects of linkage disequilibrium on the transcriptome. *Rna, 18* (1), 77-87. doi:10.1261/rna.029900.111

93 Millonig, G., Muckenthaler, M. U., & Mueller, S. (2010). Hyperferritinaemia-cataract syndrome: worldwide mutations and phenotype of an increasingly diagnosed genetic disorder. *Hum Genomics, 4* (4), 250-262. doi:10.1186/1479-7364-4-4-250

94 Celma Nos, F., Hernandez, G., Ferrer-Cortes, X., Hernandez-Rodriguez, I., Navarro-Almenzar, B., Fuster, J. L., et al. (2021). Hereditary Hyperferritinemia Cataract Syndrome: Ferritin L Gene and Physiopathology behind the Disease-Report of New Cases. *Int J Mol Sci, 22* (11). doi:10.3390/ijms22115451

95 Modell, B., & Darlison, M. (2008). Global epidemiology of haemoglobin disorders and derived service indicators. *Bull World Health Organ, 86* (6), 480-487. doi:10.2471/blt.06.036673

96 Shriner, D., & Rotimi, C. N. (2018). Whole-Genome-Sequence-Based Haplotypes Reveal Single Origin of the Sickle Allele during the Holocene Wet Phase. *Am J Hum Genet, 102* (4), 547-556. doi:10.1016/j.ajhg.2018.02.003

97 Piccin, A., Murphy, C., Eakins, E., Rondinelli, M. B., Daves, M., Vecchiato, C., et al. (2019). Insight into the complex pathophysiology of sickle cell anaemia and possible treatment. *Eur J Haematol, 102* (4), 319-330. doi:10.1111/ejh.13212

98 Stiller, C. A., & Parkin, D. M. (1996). Geographic and ethnic variations in the incidence of childhood cancer. *Br Med Bull, 52* (4), 682-703. doi:10.1093/oxfordjournals.bmb.a011577

99 Zheng, J., Huang, X., Tan, W., Yu, D., Du, Z., Chang, J., et al. (2016). Pancreatic cancer risk variant in LINC00673 creates a miR-1231 binding site and interferes with PTPN11 degradation. *Nat Genet, 48* (7), 747-757. doi:10.1038/ng.3568

100 Pan, W., Zhou, L., Ge, M., Zhang, B., Yang, X., Xiong, X., et al. (2016). Whole exome sequencing identifies lncRNA GAS8-AS1 and LPAR4 as novel papillary thyroid carcinoma driver alternations. *Hum Mol Genet, 25* (9), 1875-1884. doi:10.1093/hmg/ddw056

101 Weiskopf, D., Schmitz, K. S., Raadsen, M. P., Grifoni, A., Okba, N. M. A., Endeman, H., et al. (2020). Phenotype and kinetics of SARS-CoV-2-specific T cells in COVID-19 patients with acute respiratory distress syndrome. *Sci Immunol, 5* (48). doi:10.1126/sciimmunol.abd2071

102 Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., et al. (2020). Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA, 323* (11), 1061-1069. doi:10.1001/jama.2020.1585

103 Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H. D., Dingens, A. S., et al. (2020). Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell, 182* (5), 1295-1310 e1220. doi:10.1016/j.cell.2020.08.012

104 Makowski, L., Olson-Sidford, W., & J, W. W. (2021). Biological and Clinical Consequences of Integrin Binding via a Rogue RGD Motif in the SARS CoV-2 Spike Protein. *Viruses, 13* (2). doi:10.3390/v13020146

105 Rizwan, T., Kothidar, A., Meghwani, H., Sharma, V., Shobhawat, R., Saini, R., et al. (2021). Comparative analysis of SARS-CoV-2 envelope viroporin mutations from COVID-19 deceased and surviving patients revealed implications on its ion-channel activities and correlation with patient mortality. *J Biomol Struct Dyn* , 1-16. doi:10.1080/07391102.2021.1944319

106 Lakshminarasimhan, M., Maldonado, M. T., Zhou, W., Fink, A. L., & Wilson, M. A. (2008). Structural impact of three Parkinsonism-associated missense mutations on human DJ-1. *Biochemistry, 47* (5), 1381-1392. doi:10.1021/bi701189c

107 Gorner, K., Holtorf, E., Waak, J., Pham, T. T., Vogt-Weisenhorn, D. M., Wurst, W., et al. (2007). Structural determinants of the C-terminal helix-kink-helix motif essential for protein stability and survival promoting activity of DJ-1. *J Biol Chem, 282* (18), 13680-13691. doi:10.1074/jbc.M609821200

108 Bonifati, V., Rizzu, P., van Baren, M. J., Schaap, O., Breedveld, G. J., Krieger, E., et al. (2003). Mutations in the DJ-1 gene associated with autosomal recessive early-onset parkinsonism. *Science, 299* (5604), 256-259. doi:10.1126/science.1077209

109 Hardiman, O., Al-Chalabi, A., Chio, A., Corr, E. M., Logroscino, G., Robberecht, W., et al. (2017). Amyotrophic lateral sclerosis. *Nat Rev Dis Primers, 3* , 17071. doi:10.1038/nrdp.2017.71

110 Cardoso, R. M., Thayer, M. M., DiDonato, M., Lo, T. P., Bruns, C. K., Getzoff, E. D., et al. (2002). Insights into Lou Gehrig's disease from the structure and instability of the A4V mutant of human Cu,Zn superoxide dismutase. *J Mol Biol, 324* (2), 247-256. doi:10.1016/s0022-2836(02)01090-2

111 Bruijn, L. I., Houseweart, M. K., Kato, S., Anderson, K. L., Anderson, S. D., Ohama, E., et al. (1998). Aggregation and motor neuron toxicity of an ALS-linked SOD1 mutant independent from wild-type SOD1.*Science, 281* (5384), 1851-1854. doi:10.1126/science.281.5384.1851

112 Bhattacharya, R., Rose, P. W., Burley, S. K., & Prlic, A. (2017). Impact of genetic variation on three dimensional structure and function of proteins. *PLoS One, 12* (3), e0171355. doi:10.1371/journal.pone.0171355

113 Saito, M., Ogasawara, M., Inaba, Y., Osawa, Y., Nishioka, M., Yamauchi, S., et al. (2022). Successful treatment of congenital myasthenic syndrome caused by a novel compound heterozygous variant in RAPSN. *Brain Dev, 44* (1), 50-55. doi:10.1016/j.braindev.2021.09.001

114 Estephan, E. P., Zambon, A. A., Thompson, R., Polavarapu, K., Jomaa, D., Topf, A., et al. (2022). Congenital myasthenic syndrome: Correlation between clinical features and molecular diagnosis. *Eur J Neurol, 29* (3), 833-842. doi:10.1111/ene.15173

115 Sun, B., He, Z. Q., Li, Y. R., Bai, J. M., Wang, H. R., Wang, H. F., et al. (2022). Screening for SH3TC2 variants in Charcot-Marie-Tooth disease in a cohort of Chinese patients. *Acta Neurol Belg, 122* (5), 1169-1175. doi:10.1007/s13760-021-01605-5

116 Xie, W., Nangle, L. A., Zhang, W., Schimmel, P., & Yang, X. L. (2007). Long-range structural effects of a Charcot-Marie-Tooth disease-causing mutation in human glycyl-tRNA synthetase. *Proc Natl Acad Sci U S A, 104* (24), 9976-9981. doi:10.1073/pnas.0703908104

117 Gainza, P., Sverrisson, F., Monti, F., Rodola, E., Boscaini, D., Bronstein, M. M., et al. (2020). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat Methods, 17* (2), 184-192. doi:10.1038/s41592-019-0666-6

118 Zhou, J. (2022). Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nat Genet, 54* (5), 725-734. doi:10.1038/s41588-022-01065-4

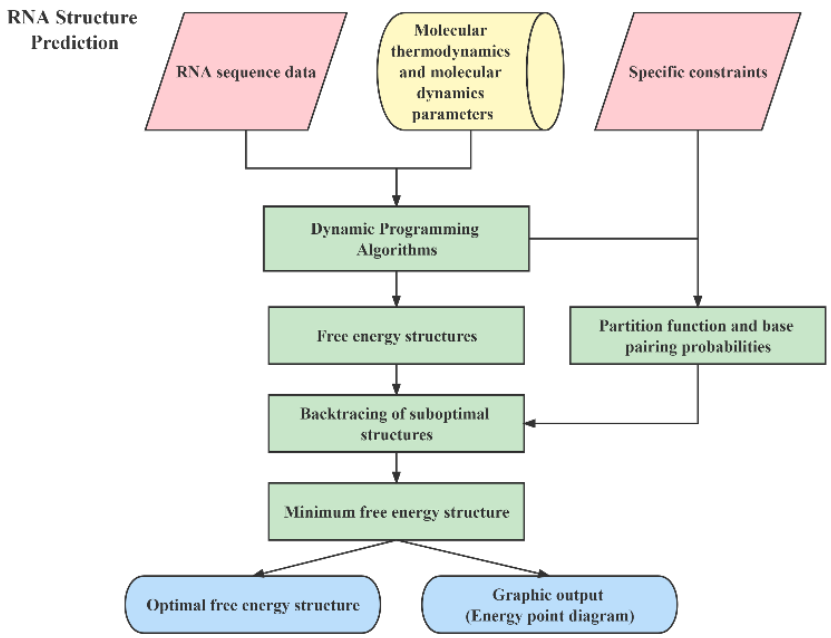

Figure 1. RNA structure prediction method flow (based on dynamic programming algorithm).

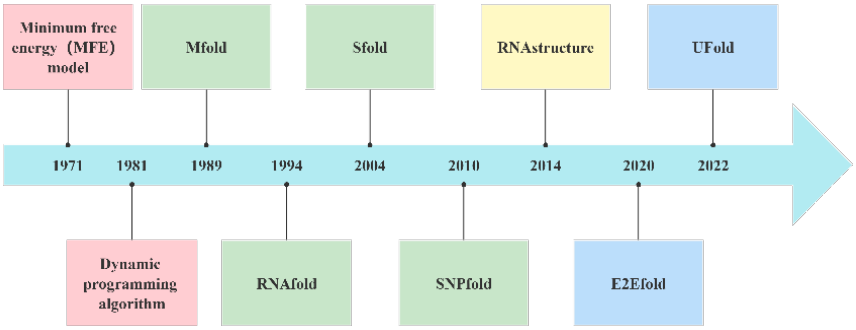

Figure 2. The development history of RNA structure prediction method.

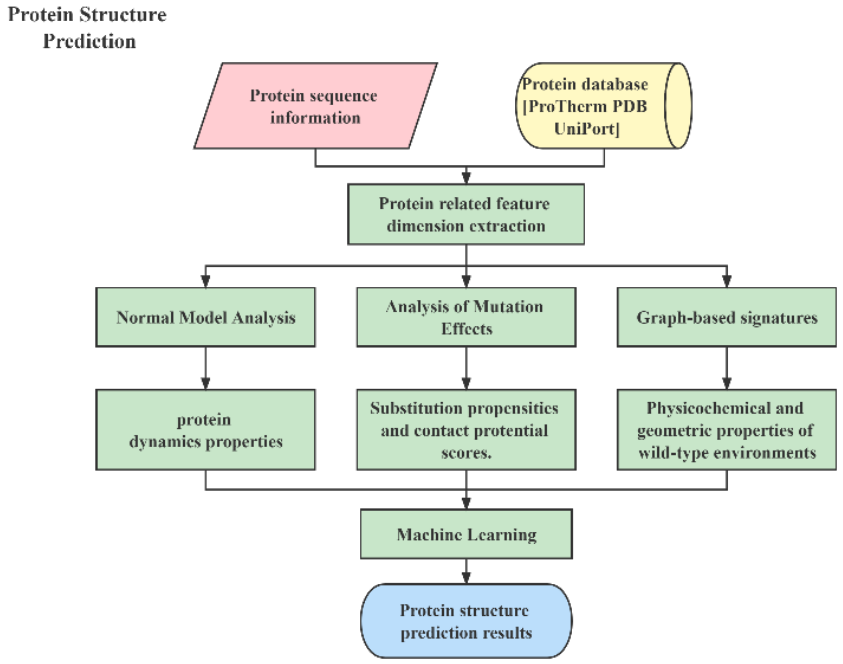

Figure 3. Protein structure prediction method flow.

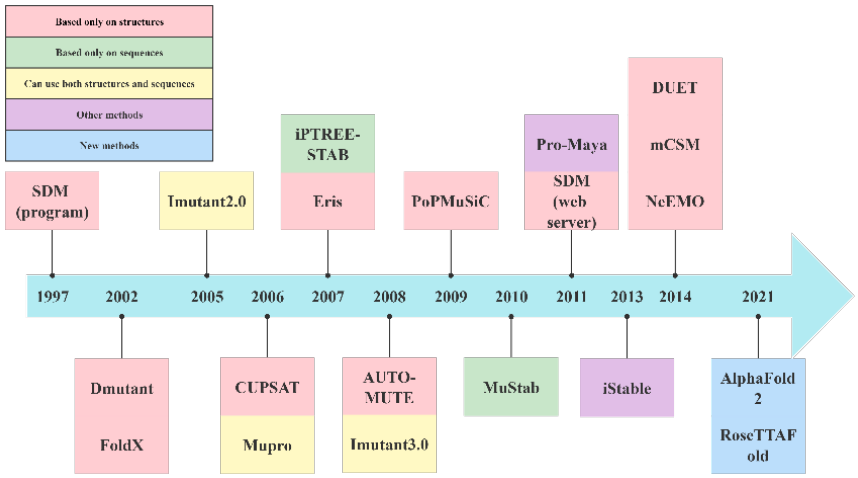

Figure 4. The development history of protein structure prediction method.

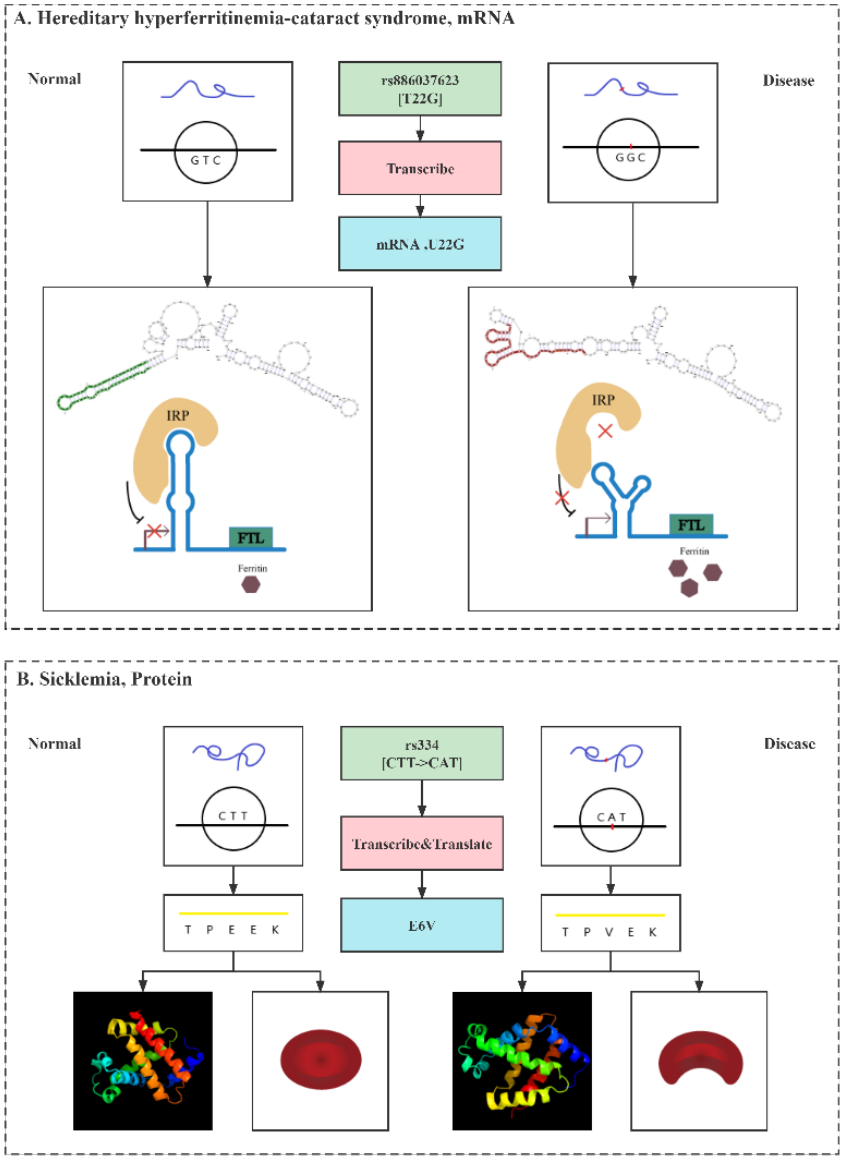

Figure 5. The influence of genetic variation on macromolecular structure in diseases.

Hosted file

`Supplementary Table 1.xlsx` available at https://authorea.com/users/624519/articles/646834-effects-of-genetic-variation-on-the-structure-of-biological-macromolecules