The future of HOPE: what can and cannot be predicted about the molecular effects of a disease causing point mutation in a protein?

Francesca Camilli[1], Annika Borrmann[1], Shima Gholizadeh[1], Tim te Beek[2], Remko Kuipers[3], Hanka Venselaar[1]

[1]CMBI, NCMLS, Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands,
[2]NBIC, Netherlands Bioinformatics Centre, Nijmegen, Netherlands,
[3]Laboratory of Systems and Synthetic Biology, Wageningen University, Wageningen, Netherlands

Depicted authors have names underlined.

Abstract

Next generation sequencing is greatly speeding up the discovery of point mutations that are causally related to disease states. Knowledge of the effects of these point mutations on the structure and function of the affected proteins is crucial for the design of follow-up experiments and diagnostic kits, and ultimately for the implementation of a cure. HOPE can automatically predict the molecular effects of point mutations. HOPE does this by massively collecting highly heterogeneous data related to the protein and the mutated residue followed by automatic reasoning that as much as possible mimics the thinking of a trained bioinformatician. We discuss HOPE and review today's possibilities and challenges in this field.

Availability: HOPE is running as a web server available at www.cmbi.ru.nl/hope/

Introduction

The development of next generation sequencing (NGS) technologies is accompanied by a series of challenges ranging from problems with storage of large amounts of data to the under-

standing of all pathways and mechanisms in an organism [1]. One of these new challenges is the analysis and prioritisation of putative disease-causing point-mutations in human genetics studies. It has been estimated that single nucleotide polymorphisms (SNPs) occur as frequently as every 100-300 bases. This implies that in an entire human genome we can potentially find 10 to 30 million SNPs [2]. The publicly available Single Nucleotide Polymorphism Database (dbSNP) nowadays contains over 30 million variations of which over 12 million are located in genes [3]. A variant is called a SNP when it occurs in at least 1% of the population. This implies that most SNPs are not directly related to a serious disease because if they were, we would all be sick. Human genomes, however, also contain many rare variants and occasionally such a rare variant causes a serious disease.

NGS is revolutionising the way human geneticists search for the causative genetic defects for disease states. In the past, extensive family tree analysis (linkage analysis) would be followed by cloning and sequencing a small region of the human genome and subsequent bioinformatics studies of the genetic variants found in this region. Nowadays, human geneticists routinely sequence the entire exome and occasionally, even the entire genome of a patient. While sequencing a human genome will become cheaper, faster, and easier in the coming years, it will remain difficult to identify which of the many observed mutations are responsible for the phenotype/disease of interest. The rate at which genomes can currently be sequenced demands for an automatic approach towards the analysis and classification of newly found variants. When hundreds of variants are detected, they must be sorted in order of likeliness that they are causative for the disease studied; this process is commonly known as prioritisation.

Prioritisation consists of two steps when variants in the exome are being analysed. First the chance must be determined that the protein is related to the phenotype studied, and second, the chance must be determined that the mutation alters the function of that protein. The protein for which the product of these two chances is highest is the best candidate for follow-up studies. The first step, determining how likely it is that a protein is related to a phenotype, is the realm of system biology. The second step, determining how likely it is that a variant alters the function of a protein, is the ultimate goal of the HOPE software, the topic of this study.

We want to know if a variation in the patient's genome is harmless or possibly disease-causing. To do this we need to compare the variation found in our patient with the 'normal' human genome. As there is no such thing as the average human, we will have to compare the variations in our patient with the variations found in a large cohort of human genomes. These variations can be found in databases, such as dbSNP [3], and are described in the OMIM database [4]. By using information extracted from such databases we can classify the variations in our patient as either 'known to be harmless', 'known to cause a disease', 'previously found mutation with unknown effect' or even as a completely new mutation that is not present in the database(s) yet. A variations that is known to be harmless (often the ones that occur frequently in a population) can be removed from our list of putative disease-causing mutations. In case the variant matches an earlier described disease-causing mutation, there is no need for further investigation because the effect of the mutation is known already. Variants that fall in the categories 'unknown' and 'completely new' are worth further investigation.

The next step is to find out whether a mutation is located in the coding sequence of a gene, or in a regulatory sequence, splice site, or otherwise functional DNA. A mutation located in a regulation site might disturb the transcription or regulation of the gene, resulting in aberrant production of the protein. In contrast, a mutation located in the protein's coding sequence is likely to affect the folding of the protein instead of its production.

To really investigate the effect of variants on the protein we need to look at its 3D-structure. Studies of the mutation in 3D, can provide insight in the effect of the mutation and lead to ideas for experiments that eventually can result in a cure for the disease studied. Unfortunately, the Protein Data Bank (PDB) provides full or partial structures for only about 20% of all human proteins, while structural information for another 20% of the human proteins can be obtained using homology modelling techniques. This leaves a 60% of the sequences without known protein 3D-structure. To find more information about these proteins we need to rely on other information sources,

such as annotations and other information in databases, conservation scores from multiple sequence alignments, and predictions based on just the sequences. It is a time-consuming task to manually collect information from all these sources, combine them, and produce a coherent idea about the effect of the studied mutation. Not every (bio)medical researcher has the tools and the experience to work with bioinformatics databases, servers, and programs. More important is the fact that it is simply impossible to manually analyse every variant in the list that results from the NGS-run. An automatic approach is required.

A series of Web servers exist that can aid with the analysis of the effects of point mutations on a proteins structure and function. Table 1 lists many of these servers together with their present internet locations.

We believe that the analysis of disease related mutations should first of all include all smart ideas in the software listed in Table 1 but, additionally, should be open and extendible so that new ideas, new concepts, new data, etc., can be incorporated quickly. We also believe that the output of a mutation analysis server should be readable by life scientists, and not only by trained bioinformaticians. We therefore developed HOPE; a fully automatic program that can collect and combine all information available for a protein (including building a homology model when required) and produces a life scientist understandable report of the mutation at hand [15].

HOPE collects information from a wide range of information sources including calculations on the 3D-coordinates of the protein using WHAT IF Web services [16,17], sequence annotations from the UniProt database [18], conservation scores from HSSP [19], and predictions by a series of Distributed Annotation System (DAS) services [20]. When possible, homology models are built with YASARA [21]. Data is stored in a database and combined in a decision scheme to identify the effects of a mutation on the protein's 3D structure and its function. The decision scheme ensures that the most reliable source of information is used for the report, being first the 3D-structure, followed by the annotations in UniProt, that in turn

Table 1. Internet based web servers that can aid with the prediction of the effects of point mutations on a proteins structure and/or function. Left hand column: name of facility, reference, and URL. Right hand column: very short description of main feature. MSA (Multiple Sequence Alignment), SVM (Support Vector Machine), PSIC (Position Specific Independent Counts), GO (Gene Ontology).

Server + URL	**Main Feature**
SIFT [5] sift.jcvi.org/	Gives one score for tolerated or not, without explanation, based on MSA.
PolyPhen [6] genetics.bwh.harvard.edu/pph/	Gives damaging or not, uses annotations, info in PDB file, MSA, and PSIC [7] scores.
PolyPhen 2 [8] genetics.bwh.harvard.edu/pph2/	As PolyPhen but with simpler and better explained output. More visualisation options.
SNPs3D [9] www.snps3d.org/	Pre-calculated 3D-effects on known protein structures; visualization using Chime.
SNAP [10] rostlab.org/services/snap/	Gives neutral or non-neutral. Uses sequence annotations, predictions on sequence, and MSA.
Panther [11] www.pantherdb.org/tools/csnpScoreForm.jsp	Gives a score for deleterious or neutral based on MSA.
PhD-SNP [12] gpcr.biocomp.unibo.it/cgi/predictors/PhD-SNP/PhD-SNP.cgi	Predicts neutral/disease based on MSA using SVM.
PMut [13] mmb2.pcb.ub.es:8080/PMut/	Uses sequence based information and predictions but not structures. A set of pre-calculated mutations on a reference PDB set is also available.
SNPS&GO [14] snps-and-go.biocomp.unibo.it/snps-and-go/	Mainly uses GO terms to indicate disease versus neutral.

are followed by sequence-based predictions. The user can submit his/her sequence and mutation of interest via the web-interface. The report will be shown at the same website and is illustrated with figures and animations showing the effects of the mutation.

While HOPE has been shown to often provide very accurate descriptions of the expected effects of mutations, it most certainly also makes the occasional error and it has limitations in terms of which bioinformatics aspects of mutant analyses it can address. We validated the software by repeating a large number of mutation analyses we performed manually in recent years and by analysing mutations described recently in articles published in high quality journals. In these articles the authors describe their analysis of the structure of the protein and the structural and/or functional effects of the mutation(s). This extensive study revealed a few HOPE-improvements that we have already implemented, and potential improvements that for a series of reasons cannot be implemented yet. Surprisingly, it also revealed a number of instances in which the authors of peer-reviewed articles in highly respected journals made errors in the bioinformatics underlying their conclusions regarding the molecular effects of the mutation studied. We believe that HOPE can help the human genetics community by providing a 'second opinion' to referees, and perhaps also to the human geneticists publishing their results. However, it should be kept in mind that HOPE is software and thus equally fallible as a human being.

Method

In recent years we collaborated in numerous human genetics projects, performing the mutation analyses, providing insight in the structural effects of mutations and, in some cases, we could also provide suggestions for new experiments. Often these studies required building a homology model, but occasionally structure information could not be obtained so that the use of sequence-based prediction servers was required to obtain all information available for the protein. Here we use these examples to validate HOPE. To extend the validation beyond our own projects that, after-all, guided the design of HOPE, we decided to test HOPE using projects that were recently described in well-known journals, such as the American Journal of Human Genetics, Nature Genetics, and Human Mutation. These journals are known to contain many articles about disease-causing mutations and their structural effects. We selected a list of mutations and performed the analyses both manually, using YASARA for model building and visualization, and automatically, using HOPE for modelling and analyses. By comparing the results we obtain an overview of the strong and weak points of HOPE, and of features that can be improved or added to the system.

Results and discussion

The full results of the analyses can be found at the HOPE website[1] and are summarized in Table 2. We classified HOPE's result as 'good' when the HOPE report contained a clear and correct description of the effect of the mutation on the 3D-structure and/or function of the protein. A result received the classification 'OK' when it contained most but not all crucial points about the mutations and no erroneous remarks were found in the report. Of course, we want the report to be as complete as possible but it will take years before we can include every possible information-source. Therefore, possible points for improvement are mentioned in the last column of the Table. Results that were fully correct and did no teach us anything about possible HOPE improvements are not listed in the Table but are available at the website. Since March 2010 users from all over the world have visited the website more than 1600 times.

In these in-house studies we compared the manual and automatic analyses of 79 mutations in 26 proteins. The number of mutations per

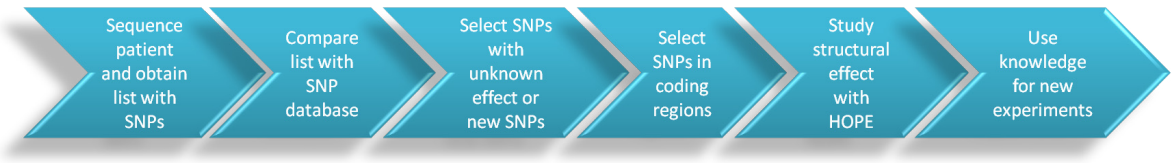

Fig 1. This figure shows how HOPE fits in the pipeline for SNP analysis.

1 www.cmbi.ru.nl/~hvensela/HOPEresults/

Table 2. Mutation analyses on in-house projects. Mutations shown in bold were explained using an experimentally solved structure, mutations in grey indicate the ones for which neither a structure nor a modelling template was available. The remaining mutations were explained using a homology model.

Protein (UniProt accession code) and reference	Mutation	HOPE's performance/points for improvement
HFE _ human PMID:18042412	**H63D** **G93R** **I105T** **L183P** **C282Y**	Good, using the complex structure and indirect interactions with other molecules could improve the results
EHMT1 _ Human PMID:19264732	**C1042Y** **R1166W**	Good, indirect dimer-interactions and a 'does the rotamer fit'-option could improve the results[1]
NDP _ Human PMID:20340138	C55R G67E G67R F89L S92P P98L K104N	As good as possible without a model, C55 is predicted buried which is not underlined by a low-res model, missing literature info for F89, the low resolution model also has info about putative cysteine bonds in the vicinity of S92 and P98
TOMT _ Human PMID:18953341	R81Q W105R E110K	Good, could benefit from a ligand contact analysis that also includes neighbouring residues
PO3F4 _ Human PMID:19671658	R329P	Good, analysis of contacts made with neighbouring residues could improve the answer
TMPS6 _ HUMAN	C702F R774C	Good, misses a possible new cysteine-bond for R774C
NDUF3 _ Human PMID:19463981	M1T G77R R122P	Good, but almost no info for M1T, result for R122P could benefit from knowledge about active site locations[1]
SEC63 _ Human PMID:20095989	I120T / D168H R217C / R267S Q375P / W651G D675E	HOPE's model differs from the manually built model and results in different accessibility-scores for D168. Others are OK
GLU2B _ Human PMID:20095989	R139H K155R M175V T261S R281W E381K	As good as possible without a model. A helical wheel predictor to identify the hydrophobic side of the helix could improve the result for E381K
KCNA1 _ Human PMID:19903818	N255D	OK, could benefit from literature information about the location of the voltage sensor.
TRPM6 _ Human PMID:18490453	G1955A	OK, could benefit from information about the glycine-rich motif.
NDUV1 _ Human	L53P/ P122L Y204C / C206G A211V / R257Q A341V / T423M	Good, results for P122L could be improved using the complete complex for analysis and information about residue stacking for R257.
NDUS2 _ Human	F84L E104G R228Q P229Q S413P D446N	OK, slight difference between the manual and automatic model, could benefit from using the complex for R228Q and information about the location of the membrane.
NDUS8 _ Human	P79L R94C R102H	Good, R94 could benefit from using the complex structure.

1 The 'does the mutated residue fit' option has been implemented as a result of this validation experiment. Additional WHAT IF options that have been added to HOPE are 'does the mutated residue make hydrogen bonds', 'does the mutated residue make a salt bridge', and 'does the mutated residue influence the shape of a cavity'.

Table 2 (cont.)		
PCD15 _ Human PMID:18719945	R134G D178G G262D	Good, result for R134G could improve by using a model that covers more domains.
SMAD3 _ Human PMID:21217753	**T261I** **R287W**	Good, result for T261I could improve using the complex structure
TLR2 _ Human	T411I R579H P631H R753Q	OK, could be improved using long distance relations and neighbour analysis in the complex.
ACAD9 _ Human PMID:20816094	E413K R518H	Good, result for R518 could benefit from long distance relations.

protein ranged from one to eight, and the protein length ranged from a small single-domain protein of 133 residues to a large multi-domain protein of 2022 residues. Nine mutations could be explained using the experimentally solved structure of the protein or protein-domain while 53 mutations could be explained using a homology model. In some cases we had to build multiple models for a single protein because these domains were only available as separate templates. For 17 mutation studies no solved structure or template could be identified and therefore our analyses had to rely on sequence based predictions and annotations. HOPE uses a very safe homology modelling threshold to make sure that the models are build only when a good template is available. Consequently, HOPE does not identify a template for NDP _ Human. However, the cysteine pattern in the sequence indicates that the protein adopts a cysteine-knot fold [22]. We were able to manually build and use a homology model for NDP. This is the only project in which we used different information sources for the manual and automatic approaches.

To extend HOPE's validation beyond our in-house projects we also analysed mutations that were reported in the literature. The selection criteria used to select test-cases from the literature included: one or more mutations were found to cause a disease, a description of the structural effects of the mutation(s) given in the article, and, if possible, a description of the model building process. The results are summarised in Table 3. Again, we classified the results as good when HOPE was able to give a clear report that agrees with our manual analysis and that is as complete as possible, while 'OK' was used for correct but incomplete cases. As in Table 1, results are not listed here if they would not teach us anything about potential HOPE improvements.

We analysed 66 mutations in 32 proteins of which 27 could be explained using the experimentally solved protein structure, 32 could be explained using a homology model. For the remaining 7 mutations we used other information sources.

Sometimes, the protein of interest was solved multiple times under different conditions. This means that HOPE had to choose which of these PDB-files to use for the analyses and/or model building. A decision schedule in HOPE will decide which template/structure to use based on the length of the aligned sequence, the percentage identity, the resolution of the solved structure, and of course on the necessity that the mutated residue must be part of the model. It sometimes happens that the authors of the article decided to use a different PDB-file. In case of the KFL1-project, for instance, this makes sense because a better template for modelling was not solved until after the article was published. In other cases the authors used experimental knowledge that only they could have to decide on a certain PDB-file as template because it contains the protein in a certain state such as active/inactive, open/closed, or bound to a certain ligand. As a result, some of HOPE's analyses were performed using a different PDB-file but in most cases this did not affect the outcome of the analyses. The choice of PDB file or modelling template indeed is a point of concern; Bywater recently worded this problem nicely [23]. Should we model the inactive state or the active state? Actually, we should probably model both states because if a mutation influences either one of the two, it will already influence the protein's function.

Table 3. Mutation analysis on previously reported mutations. Columns and fonts as for Table 2.

Protein + reference	Mutation	HOPE's performance
CHSTE _ Human PMID: 20004762	**R135P** **L137G** **R231P** **Y293C**	Good, could be improved using motif information from literature.
DPM3 _ Human PMID: 19576565	**L85S**	OK, could be improved using dimer-structure and a coiled-coil predictor.
CSKP _ Human PMID: 19200522	**R28L**	As good as possible, could be improved using splice-site analysis.
ALR _ Human PMID: 19409522	R194H	Good (almost no info in article at all)
ACTA _ Human PMID: 19409525	R39H R118C R149C R185Q R258C R258H	OK, could benefit from annotations about the location of the nucleic binding cleft, or a service that calculates this.
EMG1/NEP _ human PMID: 19463982	D86G	Good, could benefit from analysis of contacts made by neighbouring residues.
TRPV4 _ Human PMID: 19232556	D333G	As good as possible, ANK-repeat not annotated, protein becomes more active.
LRCC50 _ Human PMID: 19944405	L175R	Good, could be improved using information from literature.
RENI _ Human PMID: 21036942	**D38N** **S69Y**	Good, could be improved using a more extensive analysis of the contact residues (S69Y).
SPSY _ Human (SMS) PMID: 20556796	**G56S** **V132G** **I150T**	Good, improved the conclusions drawn by the authors.
KLF1 _ Human PMID: 21055716	E325K (better modelling template now)	OK, but HOPE misses possible new interactions formed after mutation.
FXRD1 _ human PMID: 20858599	R325W	Good, even though no model was built (template identity does not exceed HOPE's safe modelling threshold).
PSB8 _ Human PMID: 21129723	T75M	OK, could benefit from better annotation of the active site residues
PPA5 _ Human (ACP5/TRAP) PMID: 21217755	**T89I** **G215R** **D241N** **M264K**	Good, finds points not mentioned in the article
PRPS1 _ Human PMID: 20021999	**D65N** **A87T** **I290T** **G306R**	OK, could be improved using the complete hexameric biological unit (not in PDB).
ABCAD _ Human PMID: 19944402	T4031A	Good, could be improved using information about the motifs in literature.
DCTN1 _ Human PMID: 19136952	**G71A**	Good, could benefit from motif information found in literature.
PGDH _ Human PMID: 18500342	**A140P**	Could benefit dramatically from a better neighbour analysis.

In our test-cases HOPE did not make any dramatic mistakes. However, some of the, otherwise correct, answers can be improved by the implementation of new Web services, by a smarter

choice of modelling templates, or by the use of literature information. Table 4 shows a short summary of HOPE's strong points (green), points that will be improved in the (near) future (orange), and points that will not be improved soon (red). These points will be discussed more extensively below.

HOPE can: *collect structural information from the 3D-structure or build a homology model when required*

A protein's 3D-structure contains an enormous amount of useful information. The fact that HOPE builds and uses the homology model is one of its strong points because this doubles the percentage of human proteins for which 3D analysis is possible. The YASARA modelling script used in HOPE was one of the top-performers in the CASP 2008 and 2010 competition [24]. We have to keep in mind that every homology model represents only a prediction of the truth. However, the choice for one of the best modelling methods and the use of a safe homology-modelling threshold reduces the chance that HOPE analyses a completely wrong model.

HOPE can: *use the most reliable information source and combine them*

HOPE will always provide an answer. Even when there is hardly any information known about the protein, HOPE can still use predictions and information about the amino acids. The fact that HOPE uses structure, annotations, predictions, and conservation scores makes that HOPE gives more complete answers than most other servers that often use just one source of information.

HOPE can: *give a clear and understandable answer for everyone in the (bio)medical fields*

HOPE aims to serve a group of users in the field of life sciences that typically lack extensive bioinformatics experience. Therefore, the HOPE website and reports are as easy to use and understand as possible. Difficult bioinformatics keywords in the report are linked to our freely available online dictionary[2] that is based on WikiPedia's software[3].

HOPE will, in the near future, be able to: *choose the structure/templates for modelling*

Template selection is difficult but occasionally crucial. If the template includes an interaction partner, knowledge about disturbance of the interface can be gained. If an enzyme has an active and an inactive form, then both should be modelled and analysed as disturbing any of

Table 4. Summary of HOPE's strong and weak points.

HOPE can:	collect structural information from the 3D-structure or build a homology model when required
	use the most reliable information source and combine them (known structure/homology model, UniProt annotations, conservation scores, predictions)
	give a clear and understandable answer for everyone in the (bio)medical field
HOPE will, in the near future, be able to:	choose the structure/templates for modelling in a more intelligent way, keeping in mind that the protein might be solved in different conformations, complexed with different ligands, and/or under varying conditions
	use more information from new DAS servers or other sources
	analyse also the mutated situation and compare this to the wild-type, model or structure
	analyse long-distance relations
HOPE will not easily be able to:	use all information in the heads of the specialists all over the world
	to extract information from literature

2 www.cmbi.ru.nl/wiki/
3 http://www.mediawiki.org

the two states will disturb the function. Currently, HOPE can detect multimeric interactions in case the used PDB file contains a multimer. However, not every protein was solved in its biological assembly making it difficult to detect the interactions between the protein of interest and its partners, and transient interaction partners are only seldom co-crystallised. For example, for the mutations in SMAD3 [25] HOPE performs its analyses using the PDB file of the SMAD3 monomer (1mjs [26]) instead of the trimeric complex (1u7f [27]) because the first one has a significantly better resolution (1.91 vs 2.60 Å). Analyses of the mutation T281I could benefit from using the trimeric complex because this residue obviously makes protein-protein interactions at the trimer interface. In this example the interactions are not mentioned in the report, although HOPE does mention the possibility to form these interactions. In the future we want to improve the choice for templates/structures by incorporating biologically assemblies from the protein interaction database PISA [28] and by using smarter algorithms for structure choices.

HOPE will, in the near future, be able to: *use more information from new DAS servers or other sources*

Nowadays, we can access an ever increasing number of servers and databases that all provide useful information. The latest NAR special volume on databases [29] lists hundreds of databases that all might for one project or another contain useful information, but obviously today's technological possibilities preclude use of all these databases. We do intend to let the number of databases grow that HOPE can tap in to, but logistics and maintenance issues will limit us to dozens of databases rather than hundreds. HOPE could then use this information and would give a more detailed report including this domain information. The validation of HOPE gave us new ideas for possible structure calculations and prediction services that can be used to improve the reports. For instance, mutation R28L in the CSKP project [30] probably affects a splice site. As soon as there is a server that provides information about splice sites in protein sequences we can include this in the HOPE reports, and if nobody makes such a server, we will in due time (have to) do it ourselves. Mutation L85S in DPM3 [31] shows that a prediction of coiled-coil domains can be useful, especially since this seems to be the only information known about this position. We also found new ideas for WHAT IF calculations that would improve the results and a few of them have already been implemented in the system (like whether a residue is lining the wall of a cavity).

HOPE will, in the near future, be able to: *analyse also the mutated situation and compare this to the wild-type model or structure*

Insufficient detail is obtained when only the wild-type residue is analysed because the model of the mutant occasionally adds information. For instance, in the case of mutations in SPSY [32] we see that the mutant residue could possibly form new interactions thereby stabilising the protein structure and changing its behaviour. Currently, HOPE only finds the interactions that cause a loss of stability, not the ones that cause gain of stability. In the future we want to implement a module that looks for newly formed beneficial interactions. We already implemented, for example, a module that looks at the stabilising effects of prolines near the N-terminus of helices.

HOPE will, in the near future, be able to: *analyse long-distance relations*

All residues are in close contact with others. Mutation of one residue will thus also affect its spatial neighbours. For example, mutation R329P in PO3F4 [33] changes an arginine that forms hydrogenbonds an asparagine that binds directly to DNA. Mutation of this arginine can therefore indirectly affect DNA binding even though the residue itself was not found to contact DNA. In similar ways, mutations can affect ligand binding sites, active sites, etc., even when the residue itself is not found in such sites. To find these effects we plan to extend the HOPE modules with a neighbour-analysis module that considers all residues that make contact with the mutated residue. Analysis of distance relations that span more than two residues will remain difficult.

HOPE will not be able to: *use all information in the heads of the specialists all over the world*

Common knowledge obtained by years of experience in looking at protein structures cannot be stored in a database. In case of the R122P mutation in NDUF3 [34], experience tells us that the mutation is located at the same side where

you can usually find the active site in homologous proteins. Today there is no easy way, yet, to annotate this type of information.

HOPE will not be able to: *extract information from literature*

Unfortunately, there is lots of information in the literature that is not (yet) stored in an easy accessible database. Sometimes this can simply be solved by annotating the information in the UniProt database. For instance, the location of the voltage-sensor in KCNA1 or the G-motif in TRPM6 can easily be added to the sequence features of the UniProt-records for these proteins. Analysing the results of HOPE for almost a hundred published cases, we realised that a trained protein structure bioinformatician knows an amazingly large number of 'little facts'. Putting this all in software will require a few more years of programming artificial intelligence code. Machine learning will not be able to do this work for us, as the number of 'little facts' that need to be encoded is still very much larger than the number of well analysed disease causing human variants.

Example result

We would like to share one of our projects as an example of what can already be done by HOPE and what will need to be improved in the future. Two mutations were analysed in protein EHMT1; C1042Y and R1166W. The protein structure of the domain of interest was solved and can be found in PDB file 3hna [35].

By studying this protein structure we could see that the cysteine at position 1042 makes important interactions with one of the zinc-ions in the zinc-cluster in this protein. This cluster is probably important for stabilisation of the local structure

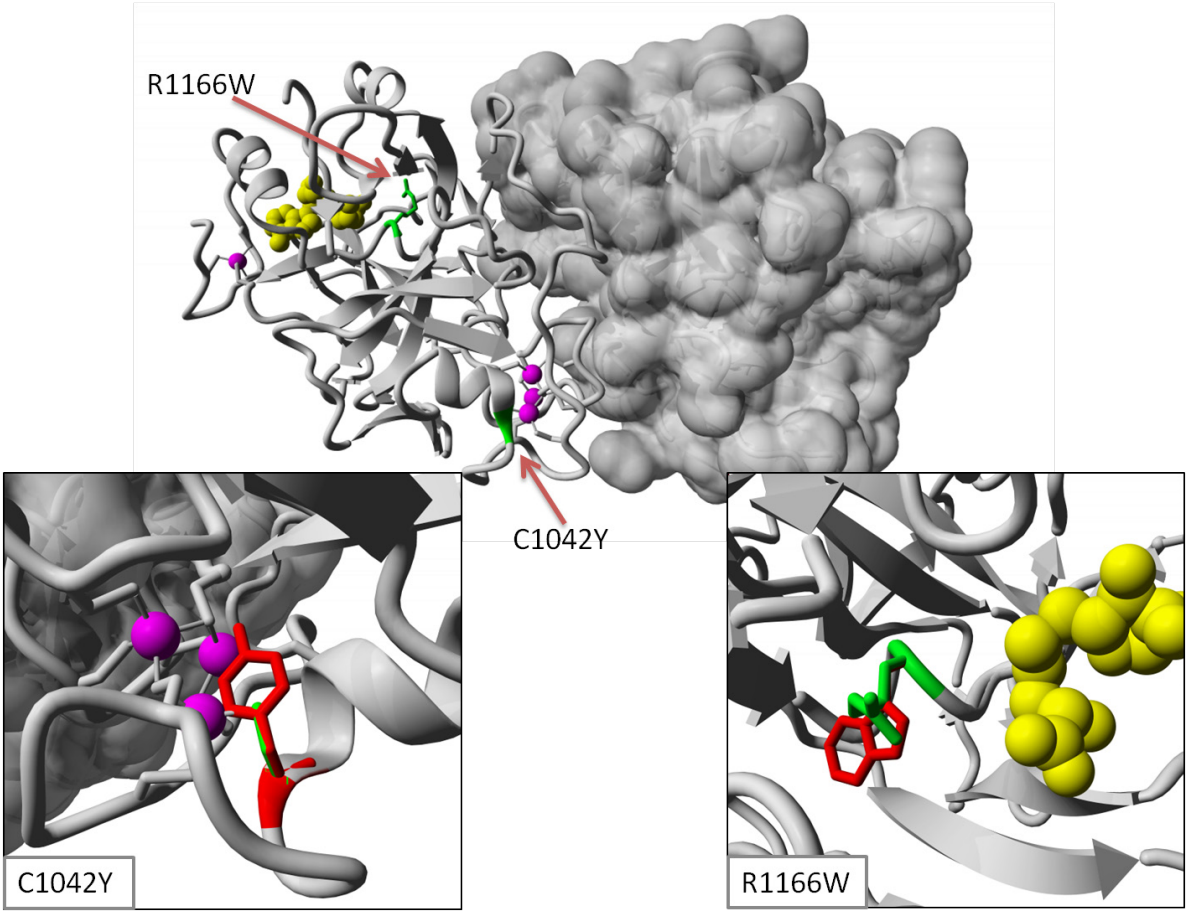

Figure 2. Overview of the mutations C1042Y and R1166W in the 3D-structure of EHMT1. The surface of only one of the monomers in the dimer is shown, the other monomer is depicted in cartoon representation. The mutated residues are colored green and indicated with red arrows. The Zinc-ions are shown as magenta balls while the ligand is shown in yellow balls. The insets show a close-up of the mutations. The side chain of the wild-type residue is now shown in green while the side chain of the mutant residue is shown in red.

element that seems needed to correctly position the loops that make interactions with the other monomer. The mutation will cause the introduction of a bigger residue which will simply not fit here and will therefore affect dimerisation. HOPE mentions the same points in its report: the interactions with the zinc-ion will be lost and the bigger residue will not fit at the same position. However, HOPE misses the fact that this could affect dimerisation because the mutated residue is not in direct contact with the other monomer. As soon as we have implemented a more extensive contact-analysis, HOPE will also be able to identify this effect.

The second mutation converts arginine 1166 into a tryptophan. In the protein structure we can see that this residue is buried and in contact with the ligand. We used a WHAT IF option to find out that no rotamer of tryptophan will ever fit at this position. The mutant will disturb the structure of the ligand binding site. HOPE also produces a report that mentions the interaction between R1166 and the ligand, that a bigger residue will probably not fit at the same position and that this will disturb interaction with the ligand. However, HOPE did not try to fit all possible rotamers of tryptophan. We are currently implementing this option.

This example shows that HOPE can already give a correct and informative answer that can be obtained easily and automatically. It also shows that there are still possibilities to improve the system and to provide even more clear and stronger answers.

We even found a few cases in which HOPE provided significantly more information than could be found in the article. For example, the authors of the ALR _ human project [36] performed a large number of experiments to find out that the mutation affects the function of the protein and might cause complex IV deficiency. HOPE mentions that the mutation is located in the ERV/ALR sulfhydryl oxidase domain and makes hydrogenbonds to FAD. The difference in size and charge will disturb this interaction which will in turn affect the function of the protein.

In this second example we show that HOPE can even improve results of mutation analyses. In their study of mutations in SPMSY causing Snyder-Robinson-Syndrome the authors describe mutation I150T. They used the experimentally solved structure of the SPMSY and found that the mutant residue threonine could make a new hydrogenbond with aspartate 222 in the hydrophobic core. We performed the same analysis but could not identify the same hydrogenbond. The minimum distance between the side chains of threonine and aspartate was found to be 4.5Å whereas a maximum distance of 3.5Å is required for hydrogenbond formation. It seems unlikely that this hydrogenbond is formed. HOPE produces a report that agrees with our manual analysis. This illustrates that HOPE can be used as engine to aid both authors and referees.

Conclusion

We have developed HOPE, a fully automatic mutant web server that can analyse the effect of point mutations on a protein's 3D-structure. We validated this server using a large number of well-described point mutations. We found that HOPE is able to give a clear and correct answer that in the majority of cases is similar to the results obtained by manual analysis. HOPE's performance depends on the information that is annotated or can be calculated from the structure. With this in mind, we think that HOPE performs very well in these projects providing clear and useful answers, even though they are not fully complete in some cases.

With the development of HOPE we have provided on small piece of the molecular puzzle: mutation analysis. It is now possible to automatically study the effects of point-mutations in the protein-coding region of the genome. In the future we can think of using HOPE to prioritise these mutations based on the probability that the mutation is disturbing the 3D structure and as such causing a disease. HOPE's analysis can then be added to the process of analysing the results from a NGS-run.

Competing interest statement

None declared

References

1. Gisel A, Bongcam-Rudloff E (2011) EMBRACE workshop "NEXT GENERATION SEQUENCING II". EMBnet.journal, 16(1):5-7.
2. The International HapMap (2003) Project. Nature, 426(6968):789-796.
3. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29(1):308-311.

4. Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA (2000) Online Mendelian Inheritance in Man (OMIM). Hum Mutat 15(1):57-61.
5. Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res, 31(13):3812-3814.
6. Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30(17):3894-3900.
7. Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, Kuznetsov EN (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. Protein Eng 12(5):387-394.
8. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods, 7(4):248-249.
9. Yue P, Melamud E, Moult J (2006) SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics 7:166.
10. Bromberg Y, Rost B (2007) SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 2007, 35(11):3823-3835.
11. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A (2003): PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 13(9):2129-2141.
12. Capriotti E, Calabrese R, Casadio R (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics 22(22):2729-2734.
13. Ferrer-Costa C, Gelpi JL, Zamakola L, Parraga I, de la Cruz X, Orozco M (2005) PMUT: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics 21(14):3176-3178.
14. Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat 30(8):1237-1244.
15. Venselaar H, Te Beek TA, Kuipers RK, Hekkelman ML, Vriend G (2010) Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. BMC Bioinformatics, 11:548.
16. Hekkelman ML, Te Beek TA, Pettifer SR, Thorne D, Attwood TK, Vriend G (2010) WIWS: a protein structure bioinformatics Web service collection. Nucleic Acids Res 38(Web Server issue):W719-723.
17. Vriend G (1990) WHAT IF: a molecular modeling and drug design program. J Mol Graph 8(1):52-56, 29.
18. Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek BE, Martin MJ, McGarvey P, Gasteiger E (2009) Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinformatics 10:136.
19. Schneider R, de Daruvar A, Sander C (1997) The HSSP database of protein structure-sequence alignments. Nucleic Acids Res 25(1):226-230.
20. Prlic A, Down TA, Kulesha E, Finn RD, Kahari A, Hubbard TJ (2007) Integrating sequence and structural biology with DAS. BMC Bioinformatics 8:333.
21. Krieger E, Koraimann G, Vriend G (2002) Increasing the precision of comparative models with YASARA NOVA--a self-parameterizing force field. Proteins 47(3):393-402.
22. Meitinger T, Meindl A, Bork P, Rost B, Sander C, Haasemann M, Murken J (1993) Molecular modelling of the Norrie disease protein predicts a cystine knot growth factor tertiary structure. Nat Genet 5(4):376-380.
23. Bywater R (2010) Solving the protein folding problems. Nature Precedings.
24. Krieger E, Joo K, Lee J, Raman S, Thompson J, Tyka M, Baker D, Karplus K (2009) Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. Proteins 77 Suppl 9:114-122.
25. van de Laar IM, Oldenburg RA, Pals G, Roos-Hesselink JW, de Graaf BM, Verhagen JM, Hoedemaekers YM, Willemsen R, Severijnen LA, Venselaar H et al. (2011) Mutations in SMAD3 cause a syndromic form of aortic aneurysms and dissections with early-onset osteoarthritis. Nat Genet, 43(2):121-126.
26. Qin BY, Lam SS, Correia JJ, Lin K (2002) Smad3 allostery links TGF-beta receptor kinase acti-

vation to transcriptional control. Genes Dev 16(15):1950-1963.

27. Chacko BM, Qin BY, Tiwari A, Shi G, Lam S, Hayward LJ, De Caestecker M, Lin K (2004) Structural basis of heteromeric smad protein assembly in TGF-beta signaling. Mol Cell 15(5):813-823.
28. Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. J Mol Biol 2007, 372(3):774-797.
29. Special database issue (2011) Nucleic Acids Res 39.
30. Piluso G, D'Amico F, Saccone V, Bismuto E, Rotundo IL, Di Domenico M, Aurino S, Schwartz CE, Neri G, Nigro V (2009) A missense mutation in CASK causes FG syndrome in an Italian family. Am J Hum Genet 84(2):162-177.
31. Lefeber DJ, Schonberger J, Morava E, Guillard M, Huyben KM, Verrijp K, Grafakou O, Evangeliou A, Preijers FW, Manta P et al. (2009) Deficiency of Dol-P-Man synthase subunit DPM3 bridges the congenital disorders of glycosylation with the dystroglycanopathies. Am J Hum Genet 85(1):76-86.
32. Zhang Z, Teng S, Wang L, Schwartz CE, Alexov E (2010) Computational analysis of missense mutations causing Snyder-Robinson syndrome. Hum Mutat, 31(9):1043-1049.
33. Lee HK, Song MH, Kang M, Lee JT, Kong KA, Choi SJ, Lee KY, Venselaar H, Vriend G, Lee WS et al. (2009) Clinical and molecular characterizations of novel POU3F4 mutations reveal that DFN3 is due to null function of POU3F4 protein. Physiol Genomics, 39(3):195-201.
34. Saada A, Vogel RO, Hoefs SJ, van den Brand MA, Wessels HJ, Willems PH, Venselaar H, Shaag A, Barghuti F, Reish O et al. (2009) Mutations in NDUFAF3 (C3ORF60), encoding an NDUFAF4 (C6ORF66)-interacting complex I assembly protein, cause fatal neonatal mitochondrial disease. Am J Hum Genet 84(6):718-727.
35. Wu H, Min J, Lunin VV, Antoshenko T, Dombrovski L, Zeng H, Allali-Hassani A, Campagna-Slater V, Vedadi M, Arrowsmith CH et al. (2010) Structural biology of human H3K9 methyltransferases. PLoS One, 5(1):e8570.
36. Di Fonzo A, Ronchi D, Lodi T, Fassone E, Tigano M, Lamperti C, Corti S, Bordoni A, Fortunato F, Nizzardo M et al. (2009) The mitochondrial disulfide relay system protein GFER is mutated in autosomal-recessive myopathy with cataract and combined respiratory-chain deficiency. Am J Hum Genet 84(5):594-604.