MUTATIONEXPLORER- A WEBSERVER FOR MUTATION OF PROTEINS AND 3D VISUALIZATION OF ENERGETIC IMPACTS

A PREPRINT

Michelle Philipp*
Leipzig University
Image and Signal Processing Group
Leipzig, Germany

Christopher W. Moth*
Vanderbilt University
Center for Structural Biology
Nashville, Tennessee, USA

Nikola Ristic*
Leipzig University
Institute for Medical Physics and Biophysics
Leipzig, Germany

Johanna K.S. Tiemann
University of Copenhagen
Linderstrøm-Lang Centre for Protein Science
Copenhagen N., Denmark
and
Novozymes A/S
Lyngby, Denmark.

Florian Seufert
Leipzig University
Institute for Medical Physics and Biophysics
Leipzig, Germany

Aleksandra Panfilova University of Copenhagen
Linderstrøm-Lang Centre for Protein Science
Copenhagen N., Denmark

Jens Meiler
Vanderbilt University
Center for Structural Biology
Nashville, Tennessee, USA
and
Leipzig University Medical School
Institute for Drug Discovery
Leipzig, Germany

Peter W. Hildebrand
Leipzig University
Institute for Medical Physics and Biophysics
Leipzig, Germany
and
Charité Universitätsmedizin Berlin
Institute of Medical Physics and Biophysics
Berlin, Germany
and
Berlin Institute of Health
Berlin, Germany

Amelie Stein
University of Copenhagen
Linderstrøm-Lang Centre for Protein Science
Copenhagen N., Denmark

Daniel Wiegreffe†
Leipzig University
Image and Signal Processing Group
Leipzig, Germany

René Staritzbichler†
Leipzig University
Institute for Medical Physics and Biophysics
Leipzig, Germany

*Authors contributed equally.
†To whom correspondence should be addressed. Email: rene.staritzbichler@medizin.uni-leipzig.de and daniel@informatik.uni-leipzig.de

ABSTRACT

The possible effects of mutations on stability and function of a protein can only be understood in the context of protein 3D structure. The MUTATIONEXPLORER webserver maps sequence changes onto protein structures and allows users to study variation by inputting sequence changes. As the user enters variants, the 3D model evolves, and estimated changes in energy are highlighted. In addition to a basic per-residue input format, MUTATIONEXPLORER can also upload an entire replacement sequence. Previously the purview of desktop applications, such an upload can back-mutate PDB structures to wildtype sequence in a single step. Another supported variation source is human single nucelotide polymorphisms (SNPs), genomic coordinates input in VCF format.

Structures are flexibly colorable, not only by energetic differences, but also by hydrophobicity, sequence conservation, or other biochemical profiling. Coloring by interface score reveals mutation impacts on binding surfaces.

MUTATIONEXPLORER strives for efficiency in user experience. For example, we have prepared 45,000 PDB depositions for instant retrieval and initial display. All modeling steps are performed by Rosetta. Visualizations leverage MDsrv/Mol*. MUTATIONEXPLORER is available at: http://proteinformatics.org/mutation_explorer/

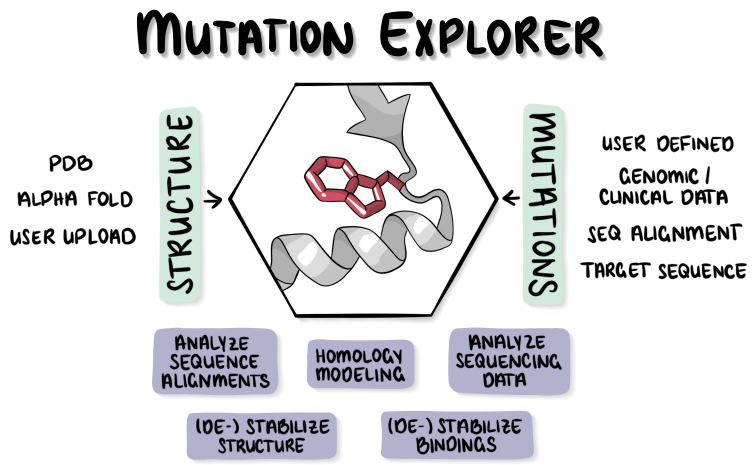

Figure 1: Graphical abstract. MUTATIONEXPLORER integrates protein 3D structure and sequence variation with Rosetta energetics and presents results as a flexible and robust website.

Mutations are essential to the persistence of life itself, enabling evolution of species in response to environmental pressures, exploited in protein engineering. Mutations can also cause challenging diseases. Throughout the biological sciences, we encounter perhaps *the* fundamental biophysical question: *How* does a mutation (or set of mutations) contribute to an observed or desired phenotype?

Specific flavors of this question include the search for explanations of phenotype variations, bacterial drug resistance, or human genetic disorders - to name only a few [1]. Many questions remain unanswered around cancer tissue, where evolution towards cell proliferation is so accelerated. In cancer, the molecular impact of mutation demands particular attention to understand disease etiology and progression.

Sequencing technologies have advanced rapidly in recent decades. As a result, over 50,000 organisms have been sequenced since 2018. Next generation sequencing (NGS) has reduced the cost of a complete human genome to less than US$1,000 [2] and today the Genome Aggregation Database (gnomAD 3.1.2) includes over 76,000 complete human genomes and over 125,000 exomes (sequenced protein-coding regions) [3]. The most frequently observed changes in proteins are missense mutations, where one amino acid is substituted by another [4]. Except for proline, all amino acids in a peptide chain share the same repeating backbone bonds atoms: -N-Cα-C-. Thus, downstream effects of missense mutations arise from biochemical changes introduced by new side chain(s). In cases of side chains which bring great energetic strain, it can be hypothesized that a protein is destabilized, perhaps to the point that it can no longer fold correctly.

In cancer, NGS increasingly informs risk, diagnosis, prognosis, and therapeutic strategies [5]. Even though NGS is not frequently employed for routine laboratory and clinical diagnostics (as only special indications are directly diagnosed by NGS), the trend towards increased patient sequencing will continue, given cost reductions in sequencing and advances in personalized medicine.

While sequencing can reveal, even predict, risks of phenotype effects or disease, as of today, sequencing alone cannot explain the molecular mechanisms by which missense mutations drive changes in protein function. Analyzing these mechanisms requires attention to protein 3D structural context at atomic resolution. Indeed, amino acids quite separated in sequence can interact closely in a folded protein.

Here we present MUTATIONEXPLORER webserver, which unites the advances in gene sequencing with advances in protein structure determination, and cements them with state-of-the-art computational methods and 3D graphics visualization. For any 3D protein model, MUTATIONEXPLORER allows both the fluid exploration of mutation sets and a more systematic treatment of mutations through sequence uploads. Supported uploads include data from sequencing experiments, multiple sequence alignments, or reference sequences. On the server's back end, Rosetta software [6] performs the mutations and calculates per-residue energies. On the web front end, mutated PDBs are displayed with the Mol* viewer [7]. Coloring scales reflect the differences in energy or hydrophobicity of the original compared to mutated proteins.

Through this flexible architecture, MUTATIONEXPLORER integrates sequence and sequence variation with the exploding availability of 3D protein structures (both experimentally determined and modelled).

The Protein Data Bank (PDB)[3] stores coordinates and related information on experimentally determined structures. However, most protein sequences are not directly represented in the PDB. Template-based modelling has long been available, and DeepMind's AlphaFold has now leveraged machine learning to model the full length of most transcripts [8]. This opens the door to new modeling approaches which take protein structure as an input for further analysis. In particular, MUTATIONEXPLORER benefits greatly from AlphaFold, as our webserver can now reveal predicted mutation energetics for a far wider array of sequence inputs.

Methods and description

Rosetta

The Rosetta software suite performs all computational modelling steps in MUTATIONEXPLORER. Rosetta debuted from David Baker's group as a tool for structure prediction. Over the years, Rosetta has been extensively reworked and extended to encompass a variety of tools for protein design, docking, and other applications [9, 6]. Rosetta incorporates both physics and knowledge-based energy potentials, the latter potentials derived from principles of statistical physics and gleaned from analysis of high resolution structures in the PDB.

When predicting structures from sequence, Rosetta executes a full Monte Carlo conformational search for the given sequence. Many protein design protocols alternate between the design stage where the backbone conformation is held fixed, while a Monte Carlo algorithm focuses on replacement of the varying side chains, and a stage where the conformation of the newly designed sequence is explored. The design step of exchanging amino acids is accompanied by an overall side chain energy optimization to allow the local environment to adjust to the new arrival. Thus, prior to the design step, a pre-minimization is required. Without the pre-minimization mathematical artifacts are more likely to arise simply because the substitution and optimization could otherwise be driven by differences between the starting structure and the minimum determined by the Rosetta energy function. The pre-minimization typically moves the coordinates by only fractions of an Ångström, well within the error of experimental detection.

For the side chain optimization Rosetta employs a rotamer library gleaned from prior exhaustive conformational analysis of the PDB. Compared to gradient-based optimization on the rugged landscape of a vast conformational space, this rotameric approach drastically reduces the search space and automatically filters out many unfavorable side chain conformations from consideration.

Rosetta writes per-residue energies into each PDB output file. MUTATIONEXPLORER writes these values into the B-factor column of the output file, allowing coloring of the entire protein molecule by its per residue energy or hydrophobicity - either in absolute values or as differences. That means, that for each coloring scheme an individual PDB file is created.

[3] https://www.wwpdb.org

Interface score One of the visualization options shows the estimated effect of the mutation on binding energies in protein-protein complexes. To obtain these estimates, the InterfaceAnalyzer Rosetta mover [10] is called through PyRosetta [11] to calculate per residue binding energies of mutated and wild type structures. The difference (MUT - WT) is calculated for every residue. InterfaceAnalyzer calculates the binding energy by scoring the input structure twice: first, as a complex, and second, after moving the two sides of the interface away from each other, exposing the interface. The Rosetta energy scores of the unbound state are then subtracted from the scores of the bound one. Both the bound (input) and unbound structures have their side chains optimized prior to scoring. Since side chain optimization is stochastic, the mover is run 3 times, and the median score is taken for every residue.

RaSP

RaSP is a new deep-learning-based tool that rapidly estimates protein stability changes. RaSP predictions strongly correlate to scores from Rosetta calculations which demand longer compute times [12]. With RaSP, MUTATIONEXPLORER presents the user with a quick initial estimation of a mutation's (de)stabilizing effect, without having to wait for the longer full minimization process. The RaSP tool consists of two linearly linked networks. A self-supervised 3D convolutional neural network that has learned representations of protein structures is followed by a fully connected neural network that maps these internal representations to Rosetta protein stability changes. RaSP is optimized for accuracy in the range [-1,7] kcal/mol, which is most relevant for revealing the loss-of-stability mutations underlying many diseases [13].

Mol* viewer

Structure predictions are visualized in MUTATIONEXPLORER with an adapted and extended version of Mol* [7], the successor of the NGL Viewer [14]. The extensions of Mol* from MDsrv [15, 16] have been incorporated into MUTATIONEXPLORER so that sequence alignments are integratively displayed near structure visualizations. During development we were in close contact with the maintainers of Mol*, and we are incorporating our extensions into the Mol* main code branch. MUTATIONEXPLORER structure visualization supports analysis in several complementary ways. After modeling the mutations, MUTATIONEXPLORER highlights structural changes with a color gradient. Alternately, through the 1D sequence alignment control, the user can easily highlight and analyze the modified regions of the structure. In order to make introduced mutations easier to examine, the alignment view from MDsrv was integrated and its capabilities were extended. In MUTATIONEXPLORER, this view now shows all the sequences of mutated versions of a protein and highlights the mutated loci in each sequence. Additionally, we developed a mutation highlighting approach which is active in the 3D structure visualization. Each mutated locus can be highlighted in a semi-transparent sphere, reducing search time and helping users to orient themselves as they rotate and translate the structure. The MUTATIONEXPLORER back end conveys not only per-residue energetic changes to the user interface, but also hydrophobicity and sequence conservation metrics. The expert can select any of these values for coloring the mutated protein. MUTATIONEXPLORER uses an adaptable color scale to visualize these differences so that the expert can tailor interpretation of impact to the various values in context of the protein of interest. When the mouse cursor is hovered over any amino acid, a panel appears which details the various computed values for the underlying residue. Clicking on a single amino acid selects it, and zooms in to the selected region. The side chain is displayed as well, to support regional visual inspection. From RaSP, every amino acid position is supported by an estimate of the global structure impact, should that position be mutated. The predicted values are shown by pressing the Ctrl key and left-clicking on the amino acid.

Sequencing data pipeline (VCF input)

For human variants provided in GRCh38 genomic coordinates [17], chromosome, position, and alt alleles may be input in Variant Call Format (VCF) [18].

We employ the Ensembl Variant Effect Predictor (VEP) [19] to analyze the VCF file and return impacted Ensembl transcript IDs and amino acid variants. The VEP often returns IDs for computationally predicted splice variants which have not been experimentally verified. We filter to retain only Ensembl transcripts which cross-reference to Swiss-Prot curated, canonical UniProt transcript IDs [20]. Finally, MUTATIONEXPLORER loads AlphaFold models [8] by UniProt ID and launches the analysis tools, as if mutations were input at protein level.

Webserver

The MUTATIONEXPLORER webserver is written in Python using the Flask framework with additional Javascript and is freely available for all users.

AlignMe

ALIGNME is a software package and webserver for detecting similarities between proteins too low to be detected on the sequence level using standard methods [21]. ALIGNME is optimized for membrane proteins, but not limited to them. ALIGNME offers a link on its result page that allows users to send their results to MUTATIONEXPLORER for 3D visualization, providing the deepest insight into an alignment.

We have made available this same Javascript interface for other sequence alignment servers to forward their results for 3D visualisation to MUTATIONEXPLORER. A link to MUTATIONEXPLORER can be easily integrated with a few lines of code. Further details are in the FAQs of MUTATIONEXPLORER.

Usage

There are two ways to define mutations for a structure when using the web page provided by the MUTATIONEXPLORER.

The Upload Structure and Mutations option is the way to go if the user wants to upload their own file or use a structure from a database such as the PDB or AlphaFold.

First, a 3D protein structure is selected in three different ways: by providing the file itself, by the definition of a PDB ID, or by the definition of an AlphaFold ID. Chains can be removed by clicking the Filter Structure button to clean up the structure. Additionally, the user can choose to minimize structures, calculate RASP predictions, and provide an email address to be notified when calculations are complete.

The mutations are defined in three different ways in the second step: by manual definition, by sequence alignment, and by definition of a target sequence. Manual definition is done by specifying which amino acid to switch at which position with a syntax like: ([chain]:[residue number][target mutation], example: "A:12G,B:134T") If a sequence alignment is provided, the user simply selects the first sequence in the alignment to match to which chain in the structure. The definition of a target sequence is done either by uploading a fasta file, by writing out the target sequence, or by providing a UniProt ID for the chain that is to be modified. When this is done, the structure is ready to be mutated and the results will be calculated.

The VCF upload: Alternately, a user can upload human SNPs, sequencing data in VCF format. For this second option, missense variants are obtained from Variant Effect Predictor output, and an AlphaFold model is selected automatically. In addition to the file upload, the user only has to decide on the minimization, the rasp calculation and whether or not to provide an e-mail address.

Via the ALIGNME server: It is also possible to import data from an ALIGNME session. After the calculation of the results on the AlignMe web server, it is possible to export the result to the MUTATIONEXPLORER. Selecting the export button on the server imports the clustal file from the calculation into MUTATIONEXPLORER. The user just needs to select which sequence from the clustal file should be the base and target sequence, and provide the actual structural file for the sequences. This can be done by providing the ID of a PDB or AlphaFold structure. Alternatively, the file can be uploaded. Additionally, the user can choose whether the structures should be minimized and whether RASP predictions should be calculated.

Navigating variants: MUTATIONEXPLORER maintains a tree-structured list of previously explored variants. A user is free to return to any model in this tree, and define further mutations (which are added to the growing tree). The workflow is sketched in Figure 2. Newly created mutations are highlighted by spheres in the 3D view, and in bright-red in the sequence view.

Coloring All scores (energy, hydrophobicity, sequence conservation and interface binding) are available for coloration. For each score type, the user can switch colors between wildtype scores, mutation scores, and their differences.

Informed by this residue-wise coloring flexibility, users can quickly hypothesize mechanistic impacts of mutation for entire molecular systems or intermolecular binding interfaces.

Navigation via sequence or 3D structure A sequence alignment window is displayed beneath the structure window. Clicking on a residue in one window highlights it in the other, and vice versa.

Visualization: Hovering reveals information about each residue. Left clicking displays side chain atoms. The "Change Visualization" button at the very bottom replaces the default sequence window with an interactive fine-tuner for color scales and overlaid spheres. The "View Alignment" button at bottom left re-displays the sequence window.

Additionally, explanations of the current color scheme are shown in an information panel. For example, sequence conservation colorings indicate the severity of an amino acid substitution using a range from white (no mutation) to light blue for mutations within amino acid groups to dark blue for mutations across groups. Sequence gaps are shown in red.

For any color scheme, a color gradient bar provides a key to the score range. Changing the bar's flanking numeric values adjusts the displayed hue range.

Mutation energy preview: Whenever RaSP calculations are included at launch (the default), the RaSP estimates of energetic effect are available via a ctrl-left click on any amino acid residue.

Continuous mutation: From the results page you are free to add new variants to the current model. Previously entered variants automatically advance to subsequent model generations.

Energy optimization: Global and thorough energy optimization before protein mutation is essential for high-quality results. Otherwise, our experience is that Rosetta will find mathematically "better" side chain conformations which are only distracting noise arising from the many small discontinuities between deposited atom coordinates and the Rosetta computational framework. Proper pre-optimization helps ensure that reported energetic differences are more directly related to the mutation itself. However, minimizing proteins, especially larger ones, is a time consuming step. MUTATIONEXPLORER offers three kind of minimizations. First, we provide a database of pre-minimized proteins from the PDB. Both backbone and side chain minimizations were performed for these. For proteins not contained in the database, the server offers either a short or a long side chain optimization, using the *fixbb* application. Alternatively, the user can upload their own minimized models. Command lines are available in the Tutorial section on the website.

A precalculated database: To build the pre-minimized database, we used a list of culled, non-redundant PDB identifiers from the PISCES server [22]. For the proteins contained in the list, we performed backbone minimization using the *relax* application of Rosetta, followed by a side chain optimization using the *fixbb* application. This yielded a database of pre-minimized structures for speeding users towards mutant exploration, currently containing 45,000 models.

Since the calculations for the minimization can be very time-consuming, the results for PDB and AlphaFold structures, which have not yet been optimized in advance, are added to the database so that the optimization does not have to be performed repeatedly. This is also done for the RaSP tool models.

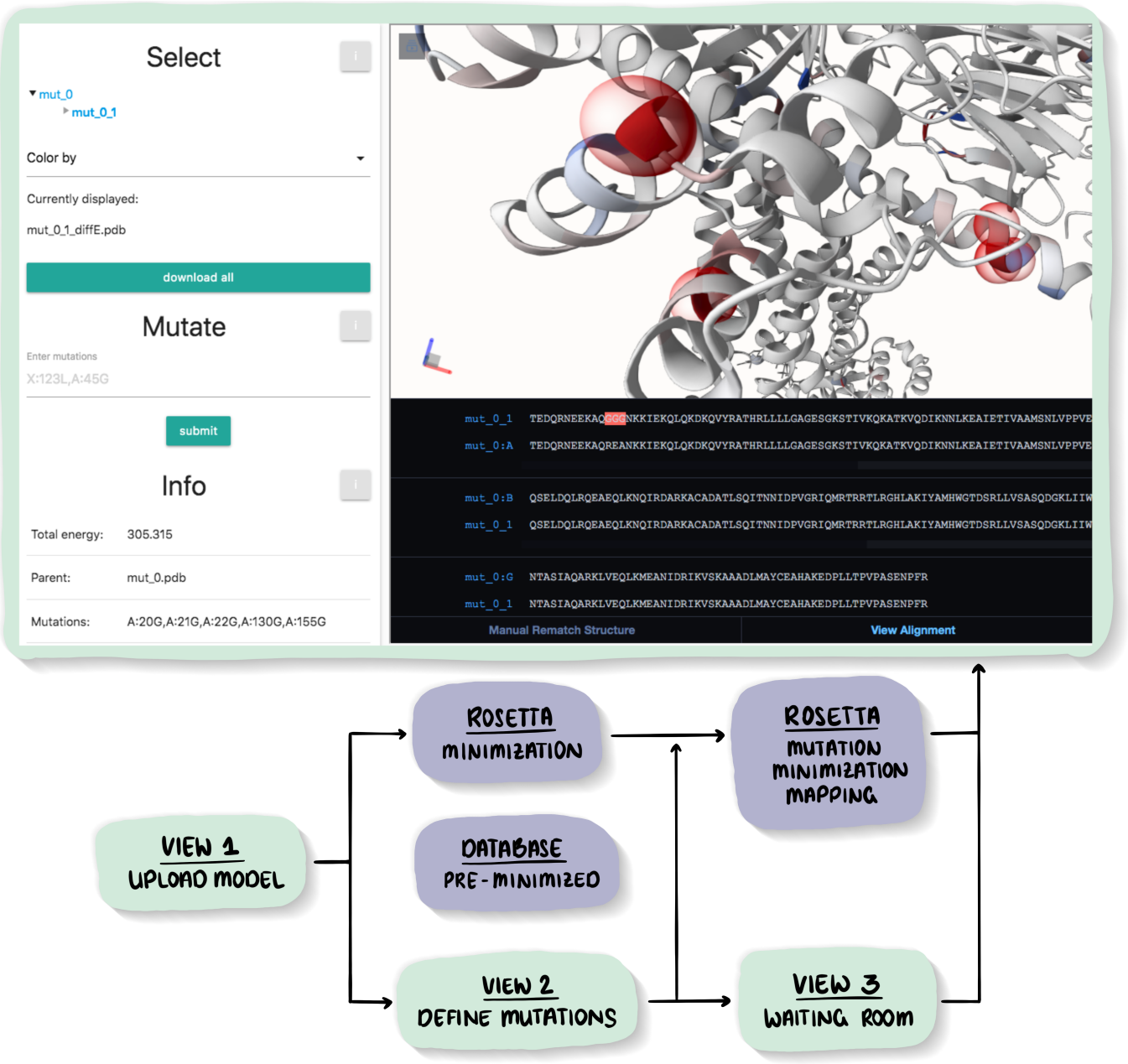

Figure 2: At the top, the result page is shown above a bubbled workflow schematic for the "upload structure and mutations" mode of the MUTATIONEXPLORER server is shown. Below is a schematic of the (internal) workflow for structure and mutation preparation. There, the schematic *user steps* (the light green VIEW bubbles) are HTML pages displayed in the browser. Shown in slate purple bubbles are the *Rosetta modeling steps* that are executed in the background on the server. The user first uploads (or selects) a protein structure or model. Second, mutations are selected. Then a ẅaiting room" page is displayed while the mutations are performed.
The result screen (top large rectangle, light green border) contains both the 3D viewer with the model of the mutated protein and below, linked to the 3D viewer, the sequence viewer shows the alignment. The same mutations are highlighted in red in both viewers. Hovering/clicking on a residue in either viewer highlights the residue in both. Additionally, a panel with details about the selected residue opens, including the energy/hydrophobicity value of the residue. Left mouse click adds an all atom licorice representation in the vicinity of the selected residue. A 'Ctrl - left mouse button' click opens a window with a fast estimate of the energy mutations at the given position will cause, calculated using RaSP [12].
The left section of the window has a selection tree of all created models, rooted at the original protein ('mut_0') and branching to all subsequent mutations. "Color by" choices include absolute or relative energies, as well as hydrophobicity. The user can continue exploring mutations by selecting a model and defining new mutations. Additionally, there is an information section, displaying the total energy of the currently displayed variant, the parent from which the current model was derived together with the mutations that were performed on the parent. Finally, the user can adjust the color scale and toggle on/off the spheres highlighting the mutated positions.

Limitations

Currently, MUTATIONEXPLORER has no potentials targeted for membrane proteins available. Especially for residues facing the membrane, a special score-function and preparation is desired. Other residues, in particular those outside of the membrane can be investigated with MUTATIONEXPLORER without limitations. Moreover, some mutations are generally challenging for Rosetta, foremost those from or to proline. Mutations from glycine may require conformational adaptation beyond the protocols used here. Mutations to glycine may cause proteins to be more flexible due to its larger torsion space. Ligands present in the PDB will be kept in rigid conformation, custom ligands will be removed. For advanced ligand handling, see [23]. MUTATIONEXPLORER cannot handle very large structures, such as CryoEM structures available only in CIF format. Partial AlphaFold models for transcripts with more than 2,700 amino acids are not supported. We also caution that for large structures our default protocols for minimizations might be insufficient. Users should definitely minimize these on their local computer before upload. For this purpose, we recommend using the corresponding tutorial on Rosetta energy minimization via the command line, which can be found on our website. The fundamental challenge is the limited sampling of individual Monte-Carlo runs. A more reliable strategy for minimization is to create an ensemble of minimized models.

Examples

The examples described below can be found on the server in the corresponding section in the main menu. In addition, there are more examples on the website under the Applications section of the tutorial, such as modeling multiple conformations of a protein and back mutation of PDB structures to wild type.

RTEL1 VCF input example

As of this writing, 248 mutations in Telomere elongation helicase 1 (RTEL1) are classified as either "Pathogenic" or "Likely pathogenic" in the Clinvar database [24]. Of these, 13 are missense mutations. No experimental structure for RTEL1 is presently deposited in the PDB. To find structural trends in pathogenicity, we uploaded the 13 Clinvar SNPs to MUTATIONEXPLORER in VCF format. Following our Rosetta based long-minimization and mutation protocols, MUTATIONEXPLORER displays the AlphaFold model for canonical transcript Q9NZ71-1. 11 residues are highlighted in the red 3D structure and sequence viewers. Two of these Clinvar variants (Chr20:63662544 A>G and Chr20:63695619 G>A) do not impact the canonical uniprot transcript, and thus are not depicted. MUTATIONEXPLORER's 11 red spheres depict all variants in obviously well-structured (high confidence) stretches of the AlphaFold model. Many of these variants involve mutations to or from proline, and MUTATIONEXPLORER generally reports large energetic differences for these, consistent with general intuition around such mutations. Indeed, several fall at secondary structure transitional points near the ends of alpha helices. Perhaps more surprisingly, our Rosetta protocol calculated quite large energetic impacts ($>$ 25 R.E.U.s) for the M492I [25] and I699M [26], both implicated in severe dyskeratosis congenita. More subtle energetic changes are also reported, reminding us that energetic analysis by itself, while necessary, is often insufficient. As example, the impact of V516L was computationally analyzed with an energetics-free model which based instead on 3D clustering patterns of pathogenic compared to benign variants [27]. In that study V516L clustered strongly with pathogenic variants, and was hypothesized to subtly disrupt DNA binding by adjacent surface residues. Link to example:
proteinformatics.org/mutation_explorer/examples/rtel1/

The GPCR autoproteolysis inducing domain

The GPCR autoproteolysis inducing (GAIN) domain is an extracellular domain of adhesion G-Protein coupled receptors, mediating receptor activation by releasing a autoproteolytically cleaved peptide from within itself [28, 29]. The GAIN domain therefore has two functions to serve - autoproteolysis and peptide release - which can be modulated by mutating the residues surrounding its cleavage site, the GPCR proteolytic site (GPS). In the rat ADGRL1 GAIN domain (PDB ID: 4DLQ), a conserved Trp residue, W804, has been demonstrated to abolish cleavage upon mutation [28]. This is reflected in substantial increases in energy for the W804S mutation in MUTATIONEXPLORER - with a similar effect when mutating another conserved Trp, and W815S - by removing the bulky side chain. In the GAIN domain, two flexible regions termed "flaps" have been shown to mediate solvent-exposure of the GPS [29]. Point-mutations disrupting interaction of i.e. Flap 2 within residues 805-810, e.g. by Y806A, also result in energy increases not limited to the flap, but also in the interacting helix, hinting at a targeted disruption of the regions interactions, likely affecting overall GAIN domain dynamics. This example demonstrates the ability of MUTATIONEXPLORER to predict the impact of point mutations on protein regions. Link to example:
proteinformatics.org/mutation_explorer/examples/agpcr_demo/

Availability

MUTATIONEXPLORER joins our series of web services available at our platform http://proteinformatics.org which we aspire to support and maintain long-term: ProteinPrompt [30], MDCiao [31], Voronoia [32], Sierra Platinum Service[33], AlignMe [21]. MUTATIONEXPLORER is freely available to all users online without any login requirement at:

http://proteinformatics.org/mutation_explorer

The MDSRV implementation on which MUTATIONEXPLORER is build is available at:

http://proteinformatics.org/mdsrv-web

The interface for alignment servers is available at:

https://github.com/starbeachlab/explorer_interface

Issues with the website can be reported at: https://github.com/starbeachlab/MutationExplorerServer/issues

Acknowledgements

We would like to express our gratitude to the team of developers and maintainers of Mol*, especially Alexander Rose, for many valuable remarks. The authors gratefully acknowledge the GWK support for funding this project by providing computing time through the Center for Information Services and HPC (ZIH) at TU Dresden.

Funding

The authors acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC 1423, project number 421152132, sub-projects Z04 (P.W.H.); as well as through the Open Access Publishing Fund of Leipzig University within the DFG program Open Access Publication Funding. This study was further funded by the Protein Interactions and Stability in Medicine and Genomics (PRISM) centre funded by the Novo Nordisk Foundation (NNF18OC0033950, to AS) and a grant from the Lundbeck Foundation (R272-2017-4528, to AS). J.M. acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG) through SFB1423 (421152132), SFB 1052 (209933838), and SPP 2363 (460865652). J.M. is supported by a Humboldt Professorship of the Alexander von Humboldt Foundation. J.M. is supported by BMBF (Federal Ministry of Education and Research) through the Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI). This work is partly supported by BMBF (Federal Ministry of Education and Research) through DAAD project 57616814 (SECAI, School of Embedded Composite AI). Work in the Meiler laboratory is further supported through the NIH (R01 HL122010, R01 DA046138, R01 AG068623, R01 CA227833, R01 LM013434, S10 OD016216, S10 OD020154, S10 OD032234).

Declarations

The authors declare no conflict of interest.

References

[1] Sebastian Köhler, Michael Gargano, Nicolas Matentzoglu, et al. The human phenotype ontology in 2021. *Nucleic Acids Research*, 49(D1):D1207–D1217, December 2020. [PubMed:33264411] [PubMed Central:PMC7778952] [doi:10.1093/nar/gkaa1043].

[2] K. Schwarze, J. Buchanan, J. M. Fermont, H. Dreau, M. W. Tilley, J. M. Taylor, P. Antoniou, S. J. L. Knight, C. Camps, M. M. Pentony, E. M. Kvikstad, S. Harris, N. Popitsch, A. T. Pagnamenta, A. Schuh, J. C. Taylor, and S. Wordsworth. The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom. *Genet Med*, 22(1):85–94, Jan 2020. [PubMed:31358947] [PubMed Central:PMC6944636] [doi:10.1038/s41436-019-0618-7].

[3] Konrad J. Karczewski, Laurent C. Francioli, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, May 2020.

[4] Konrad J. Karczewski, Laurent C. Francioli, Grace Tiao, et al. The mutational constraint spectrum quantified from variation in 141, 456 humans. *Nature*, 581(7809):434–443, May 2020.

[5] Y. Zhong, F. Xu, J. Wu, J. Schubert, and M. M. Li. Application of Next Generation Sequencing in Laboratory Medicine. *Ann Lab Med*, 41(1):25–43, Jan 2021. [PubMed:32829577] [PubMed Central:PMC7443516] [doi:10.3343/alm.2021.41.1.25].

[6] J. K. Leman, B. D. Weitzner, S. M. Lewis, et al. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat Methods*, 17(7):665–680, Jul 2020. [PubMed:32483333] [PubMed Central:PMC7603796] [doi:10.1038/s41592-020-0848-2].

[7] David Sehnal, Sebastian Bittrich, Mandar Deshpande, Radka Svobodová, Karel Berka, Václav Bazgier, Sameer Velankar, Stephen K Burley, Jaroslav Koča, Alexander S Rose, and et al. Mol* viewer: Modern web app for 3d visualization and analysis of large biomolecular structures. *Nucleic Acids Research*, 49(W1), 2021. [PubMed:33956157] [PubMed Central:PMC8262734] [doi:10.1093/nar/gkab314].

[8] J. Jumper, R. Evans, A. Pritzel, T. Green, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 08 2021. [PubMed:34265844] [PubMed Central:PMC8371605] [doi:10.1038/s41586-021-03819-2].

[9] Kim T. Simons, Rich Bonneau, Ingo Ruczinski, and David Baker. Ab initio protein structure prediction of casp iii targets using rosetta. *Proteins: Structure, Function, and Bioinformatics*, 37(S3):171–176, 1999. [PubMed:10526365] [doi:10.1002/(sici)1097-0134(1999)37:3+<171::aid-prot21>3.3.co;2-q].

[10] P. B. Stranges and B. Kuhlman. A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. *Protein Sci*, 22(1):74–82, Jan 2013. [PubMed:23139141] [PubMed Central:PMC3575862] [doi:10.1002/pro.2187].

[11] S. Chaudhury, S. Lyskov, and J. J. Gray. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*, 26(5):689–691, Mar 2010. [PubMed:20061306] [PubMed Central:PMC2828115] [doi:10.1093/bioinformatics/btq007].

[12] L. M. Blaabjerg, M. M. Kassem, L. L. Good, N. Jonsson, M. Cagiada, K. E. Johansson, W. Boomsma, A. Stein, and K. Lindorff-Larsen. Rapid protein stability prediction using deep learning representations. *Elife*, 12, May 2023. [PubMed:37184062] [PubMed Central:PMC10266766] [doi:10.7554/eLife.82593].

[13] Peng Yue, Zhaolong Li, and John Moult. Loss of protein structure stability as a major causative factor in monogenic disease. *Journal of molecular biology*, 353(2):459–473, 2005. [PubMed:16169011] [doi:10.1016/j.jmb.2005.08.020].

[14] Alexander S. Rose and Peter W. Hildebrand. NGL Viewer: a web application for molecular visualization. *Nucleic Acids Research*, 43(W1):W576–W579, 04 2015. [PubMed:25925569] [PubMed Central:PMC4489237] [doi:10.1093/nar/gkv402].

[15] Johanna K Tiemann, Ramon Guixà-González, Peter W Hildebrand, and Alexander S Rose. Mdsrv: Viewing and sharing molecular dynamics simulations on the web. *Nature Methods*, 14(12):1123–1124, 2017.

[16] Michelle Kampfrath, René Staritzbichler, Guillermo Pérez Hernández, Alexander S Rose, Johanna K S Tiemann, Gerik Scheuermann, Daniel Wiegreffe, and Peter W Hildebrand. MDsrv: visual sharing and analysis of molecular dynamics simulations. *Nucleic Acids Research*, 50(W1):W483–W489, 05 2022. [PubMed:35639717] [PubMed Central:PMC9252803] [doi:10.1093/nar/gkac398].

[17] Valerie A. Schneider, Tina Graves-Lindsay, et al. Evaluation of grch38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, 27(5):849–864, 2017. [PubMed:28396521] [PubMed Central:PMC5411779] [doi:10.1101/gr.213611.116].

[18] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin, and 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, 06 2011. [PubMed:21653522] [PubMed Central:PMC3137218] [doi:10.1093/bioinformatics/btr330].

[19] W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, A. Thormann, P. Flicek, and F. Cunningham. The Ensembl Variant Effect Predictor. *Genome Biol*, 17(1):122, Jun 2016.

[20] The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, 11 2016.

[21] René Staritzbichler, Emily Yaklich, Edoardo Sarti, Nikola Ristic, Peter W Hildebrand, and Lucy R Forrest. AlignMe: an update of the web server for alignment of membrane protein sequences. *Nucleic Acids Research*, 50(W1):W29–W35, 05 2022. [PubMed:35609986] [PubMed Central:PMC9252776] [doi:10.1093/nar/gkac391].

[22] Guoli Wang and Jr Dunbrack, Roland L. PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Research*, 33(suppl_2):W94–W98, 07 2005. [PubMed:15980589] [PubMed Central:PMC1160163] [doi:10.1093/nar/gki402].

[23] J. Mendenhall, B. P. Brown, S. Kothiwale, and J. Meiler. BCL::Conf: Improved Open-Source Knowledge-Based Conformation Sampling Using the Crystallography Open Database. *J Chem Inf Model*, 61(1):189–201, Jan 2021. [PubMed:33351632] [PubMed Central:PMC8130828] [doi:10.1021/acs.jcim.0c01140].

[24] Melissa J Landrum, Jennifer M Lee, Mark Benson, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, 46(D1):D1062–D1067, January 2018. [PubMed:29165669] [PubMed Central:PMC5753237] [doi:10.1093/nar/gkx1153].

[25] Amanda J Walne, Tom Vulliamy, Michael Kirwan, Vincent Plagnol, and Inderjeet Dokal. Constitutional mutations in RTEL1 cause severe dyskeratosis congenita. *Am. J. Hum. Genet.*, 92(3):448–453, March 2013. [PubMed:23453664] [PubMed Central:PMC3591859] [doi:10.1016/j.ajhg.2013.02.001].

[26] Tangui Le Guen, Laurent Jullien, Fabien Touzot, et al. Human RTEL1 deficiency causes Hoyeraal-Hreidarsson syndrome with short telomeres and genome instability. *Hum. Mol. Genet.*, 22(16):3239–3249, August 2013. [PubMed:23591994] [doi:10.1093/hmg/ddt178].

[27] R Michael Sivley, Jonathan H Sheehan, Jonathan A Kropski, Joy Cogan, Timothy S Blackwell, John A Phillips, William S Bush, Jens Meiler, and John A Capra. Three-dimensional spatial analysis of missense variants in RTEL1 identifies pathogenic variants in patients with familial interstitial pneumonia. *BMC Bioinformatics*, 19(1), December 2018. [PubMed:29361909] [PubMed Central:PMC5781290] [doi:10.1186/s12859-018-2010-z].

[28] Demet Araç, Antony A Boucard, Marc F Bolliger, Jenna Nguyen, S Michael Soltis, Thomas C Südhof, and Axel T Brunger. A novel evolutionarily conserved domain of cell-adhesion gpcrs mediates autoproteolysis. *The EMBO Journal*, 31(6):1364–1378, 2012. [PubMed:22333914] [PubMed Central:PMC3321182] [doi:10.1038/emboj.2012.26].

[29] Gerti Beliu, Steffen Altrichter, Ramon Guixà-González, et al. Tethered agonist exposure in intact adhesion/class b2 gpcrs through intrinsic structural flexibility of the gain domain. *Molecular Cell*, 81(5):905–921.e5, 2021. [PubMed:33497605] [doi:10.1016/j.molcel.2020.12.042].

[30] Sebastian Canzler, Markus Fischer, David Ulbricht, Nikola Ristic, Peter W Hildebrand, and René Staritzbichler. ProteinPrompt: a webserver for predicting protein–protein interactions. *Bioinformatics Advances*, 2(1), 08 2022. [PubMed:36699419] [PubMed Central:PMC9710678] [doi:10.1093/bioadv/vbac059].

[31] Guillermo Pérez-Hernández and Peter. W. Hildebrand. mdciao: Accessible analysis and visualization of molecular dynamics simulation data. *bioRxiv*, 2022. [doi:10.1101/2022.07.15.500163].

[32] Rene Staritzbichler, Nikola Ristic, Andrean Goede, Robert Preissner, and Peter W Hildebrand. Voronoia 4-ever. *Nucleic Acids Research*, 49(W1):W685–W690, 06 2021. [PubMed:34107038] [PubMed Central:PMC8265189] [doi:10.1093/nar/gkab466].

[33] Daniel Wiegreffe, Lydia Müller, Jens Steuck, Dirk Zeckzer, and Peter F Stadler. The sierra platinum service for generating peak-calls for replicated chip-seq experiments. *BMC research notes*, 11:1–5, 2018. [PubMed:30055643] [PubMed Central:PMC6064048] [doi:10.1186/s13104-018-3633-x].