MOLECULOMICS
working molecular knowledge across genomes

Human3DProteome – whole system drug discovery, drug repurposing and toxicity screening

Human3DProteome Technical Information

Platform overview

The Human3DProteome platform provides a simple web-based interface to Moleculomics' extensive protein structural modelling and affinity docking, data analysis tools, and links to biological databases.

It provides a coherent pathway taking the user from genome-level information, through protein modelling, to compound docking and analysis of potential protein-compound interactions, and on to analysis of biological context and therapeutic potential. The user can search by protein, compound, or docking characteristics, view 3D structures and docking conformations, and get reports on the likely biological relevance and *in-silico* toxicity screens.

Protein structural modelling

The platform contains the entire protein coding of the human genome, as DNA and protein amino acid sequences. High Performance Computing pipelines for homology modelling and protein threading have been applied to the whole proteome, giving high-quality 3-dimensional models of the proteins that can be used in docking and other analyses. Homology modelling gives reliable models for proteins similar to well-characterised proteins with experimentally known structures, where good templates of determined structure are available. Threading gives consistently stable models for protein sequences with lower template homology. Access to both methods, and at HPC scale, allows coverage of the entire proteome.

The content of proteome wide structural databases is highly dependent upon the quality and accuracy of the structural models of proteins within the target organisms. The precision of structural models produced is constantly improving as new protein templates are discovered based upon experimental findings and deposited in global databases of experimentally determined protein structures. The emerging data is utilised iteratively to produce refined protein structure predictions which in turn improves understanding of the interaction between proteins and compounds for the purpose of identifying new drugs and drug targets. In order for the platform to continue to benefit from these technological advances in the future, Moleculomics is able to offer provision of database updates and/or extension of the current database and platform if needed. Additional protein isoforms and variants are also being constantly added to the system.

As of summer 2017, the system contains more than 20,000 protein structures.

Homology modelling

The Moleculomics *in silico* homology modelling pipeline has been developed over a period of 12 years. Described as "world-leading", it predicts the 3D structure of proteins (drug receptors, transport proteins, ion channels, enzymes, antibodies, structural proteins etc.) directly from sequence. Multiple templates are used for each protein, and are selected based on structural similarity as well as sequence similarity.

A novel approach of structural alignment enables the test sequence to be aligned with homologous structural templates and the highest homology fit for a particular section of sequence is identified, and the corresponding secondary structure from the template is adopted. Along the length of a given sequence, a number of templates are typically used, normally 3-12 in the construction of an entire protein. Even minor changes in sequence can result in the adoption of a different template for a given section. This "template interchange" is the driver for fold changes observed in the final structures of variant proteins. Recording the changes in template coverage between wild type and variant proteins provides a rapid indication and measure of structural change.

Threading

Threading, also known as fold recognition, is another method of creating 3D structures of proteins. Instead of using sequence alignment like homology modelling, threading uses templates of known proteins that have similar folds to the sequence that has been submitted. Threading is used on proteins that do not have a high enough percentage of homologue coverage to produce a model through homology modelling. Multiple threading algorithms are used to decide the best templates to be used to model the protein sequence. This allows us to produce protein models with very high accuracy, by combining both threading and homology modelling techniques.

Structural modelling publications

The homology modelling and protein threading capabilities have been developed over many years, applied in an extensive body of research work to the modelling of hundreds of protein structures and featured in dozens of journal publications http://scholar.google.co.uk/citations?user=e5sOfe0AAAAJ&hl=en

Compound libraries

Libraries of compounds suitable for docking have been collected from several public libraries. These include all FDA approved drugs and 'experimental' drug-like compounds, via DrugBank (Law et al., 2014), which are suitable for repurposing studies; the ToxCast compound database of known toxins, which is maintained by the US Government Environmental Protection Agency, the *in vitro* testing of which has been carried out by assaying against a set of 147 human receptors and enzymes (Sipes et al., 2013); and an extensive set of molecules of suitable chemistry and biological importance from the PubChem database (Kim et al., 2016).

We have also generated a large library of ligand fragments which have been filtered via pharmacophore-based measures of toxicity. By filtering toxicophoric substructures we aim to reduce the incidence of pan-assay interference substances (PAINS) in indicated leads (Baell and Holloway, 2010). These fragments are highly suitable for fragment-to-lead in silico drug discovery. They can be similarity-searched against our library of clean lead compounds, which has been curated from the ZINC database (Sterling and Irwin, 2015) and other sources.

The system can be used for docking any user-uploaded compound against any target of interest, a user-defined panel of proteins, or the whole human proteome, making it a capable pipeline for identifying both on-target and off-target interactions.

Toxicity screening

'Panel 44' (Bowes et al., 2012) is a protein panel used for pharmacological profiling and toxicity screening. Many of the compounds in the Human3DProteome database have been run through this proprietary *in-silico*

toxicity screening process, which gives a high-throughput computational equivalent to the 'SafetyScreen44' *in vitro* assays (Eurofins Cerep-Panlabs, 2014).

Interactions are identified via several novel scoring functions trained on *in vitro* assay hit and miss data, including a step that simulates ligand displacement, and these scores are then used as components in a consensus scheme which gives a final interaction score between panel proteins and the compound of interest. If a compound scores highly across multiple receptors, it is reported as having a potential toxicity issue.

A 'traffic light' classification is available in the Human3DProteome interface to aid in initial compound filtering, and bridges to more detailed compound analysis.

Docking methodologies, analysis and interpretation

All our protein-compound interaction studies are performed using our large-scale HPC workflow, which uses a consensus approach with multiple, state of the art, docking algorithms, a combination of empirical (affinity) and force field (energy) docking. These include the well-respected and validated AutoDock Vina program (Trott and Olson, 2010).

These algorithms were chosen as they can claim the highest accuracy (>80% upon simulated docking of known co-crystallised ligands) of those docking programs that may be feasibly implemented on High Performance Computing (supercomputers). The programs also possess the great attribute of applying a high degree of ligand flexibility in the docking routines used. The ligand may be orientated and "forced" into optimal interaction with the region of protein concerned. In this simulation, the protein structure is rigid, or a molecular mechanics (MM) approach is applied locally to the residues of the active site. A higher resolution result may be attained by applying molecular dynamics (MD) approaches, which due to their high computational demands, are not feasibly applied across such large protein-ligand sets, though the protein-ligand complexes generated are a good starting point for MD work.

The raw data from these is then run through our in-house consensus analysis and re-scoring algorithms, to provide a more sophisticated interpretation of the interaction than just a raw docking score, and provide results of high biological relevance, extensively validated by reference to known *in vitro* and structural interactions (please see example workflows and case studies, below).

Access to large supercomputing facilities allows us to simulate millions of dockings across wide groups of proteins and compounds. This enables the application of sophisticated clustering approaches to categorise the interactions, and put a single protein-compound interaction in the wide context of how it compares to the breadth of interactions involving that protein or compound. This valuable contextual information cannot be obtained from a single docking in isolation.

Biological data

Biological information from a variety of databases has been cross-referenced and can be accessed through the Human3DProteome system, to inform the biological importance and connectivity of proteins and compounds, and highlight potential avenues for further research. We have developed a comprehensive *in vitro* knowledgebase from a number of leading databases, including known drug-protein interactions, linkage to metabolic and signalling pathways, diseases and therapeutic applications, tissue compartments, and interaction with the Reactome pathway knowledgebase for proteins (Fabregat et al., 2016).

Interface overview

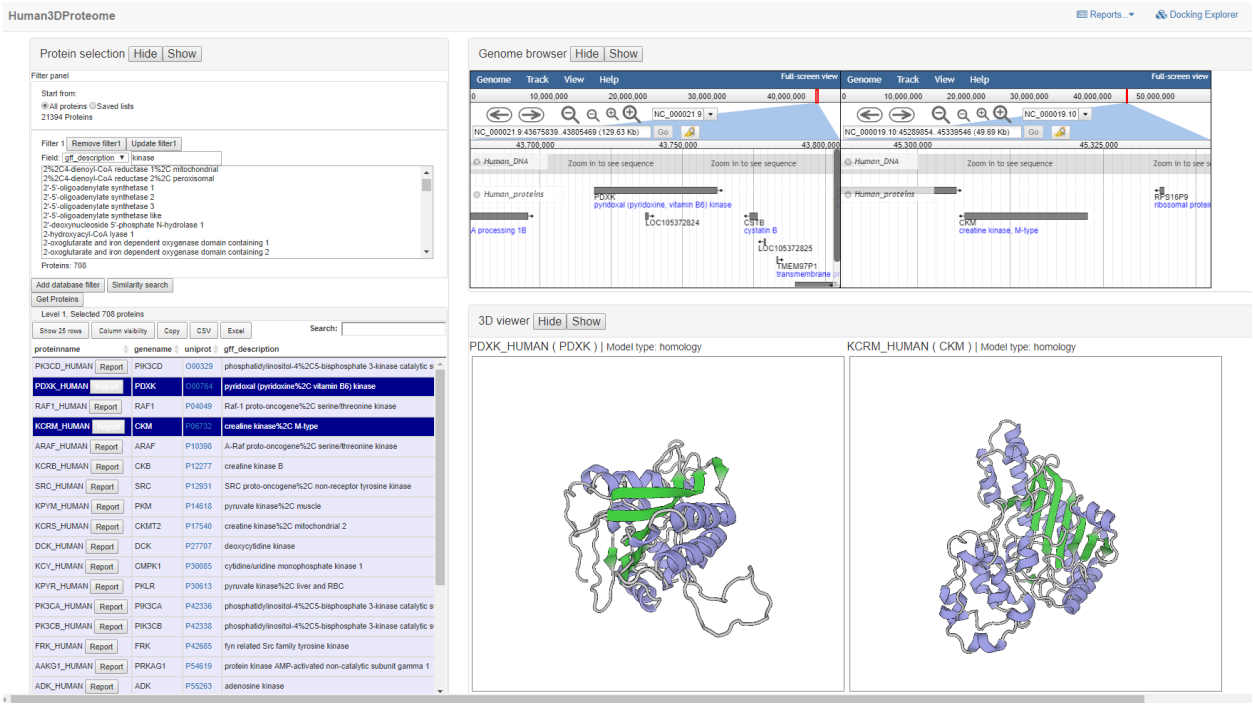
The interface provides a 'Docking Explorer' page, where users can select proteins and compounds and visualise both the genomic position of the protein and the 3D structure of it and any of all the compounds docked with it. Also provided are 'Report' pages, for looking in detail at a single protein (vs. many compounds) or compound of interest (vs. many proteins), and seeing how it relates, via dockings, to known diseases, metabolic pathways, etc.

Docking Explorer

The Human3DProteome Docking explorer page is divided into several key sections: There are three filter panels on the left, for searching for proteins, compounds, and docking results of interest, and a table of the results for each search.

At the top right is a genome browser, for viewing genome level information for each protein, right down to base-pair level. Gene annotations and descriptions are also displayed.

Underneath are several output areas, for graphical 3D views of proteins and dockings, and tabular results. Up to 4 proteins can be viewed simultaneously and compared in both the genome browser and 3D viewer.



Searching and filtering

Searching for proteins or compounds, or filtering of result sets can be performed using any field contained in the database. A graphical interface assists in creating hierarchical levels of filters to allow acquisition of highly specific results of interest. Result sets can also be stored as saved lists, available for quick access any time that particular user logs on.

Similarity searching

Several ways of extending a search by similarity are available. For compounds, Tanimoto scoring is used to select compounds similar to those specified. For proteins, both sequence similarity, based on Smith and Waterman scoring, and structural similarity, based on template use, are available.

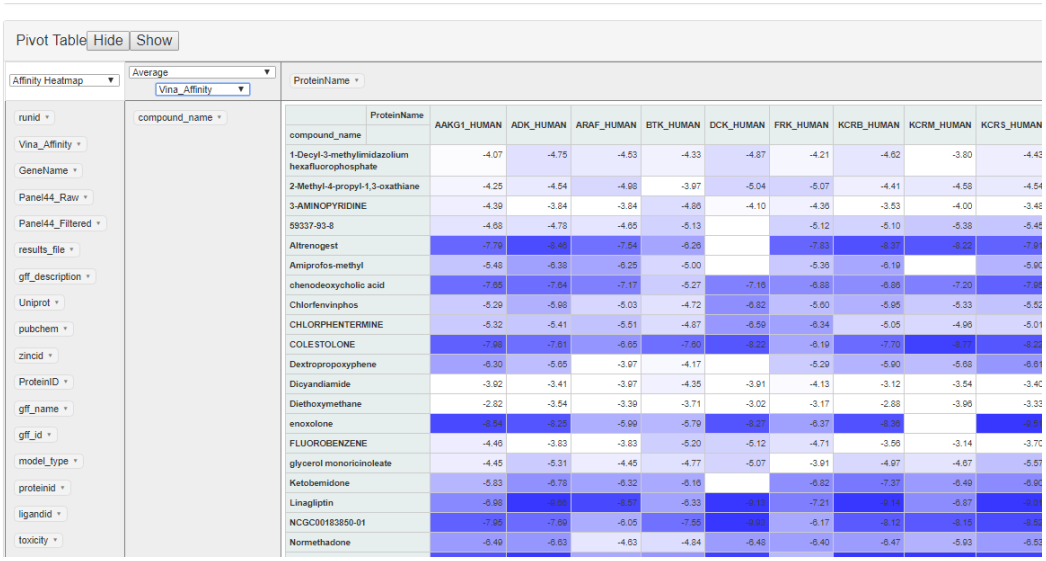
3D viewer

The 3D viewer provides an interactive visualisation of the 3D model of a protein, along with any docked compounds. The view can be rotated and zoomed, and molecules can be displayed in different styles and colouring, such as ribbons, sticks, spheres and dots. Hovering the mouse cursor over part of a protein will display the amino acid type and number. Multiple docked compounds can be displayed on each protein, and up to four proteins can be compared at once.

Data tables

For protein, compound, and docking searches, a results table is shown of raw and interpreted information from our knowledgebase. These tables interact with the other elements of the interface – clicking on a protein search result will show that gene in the genome browser and display the 3D model in the viewer.

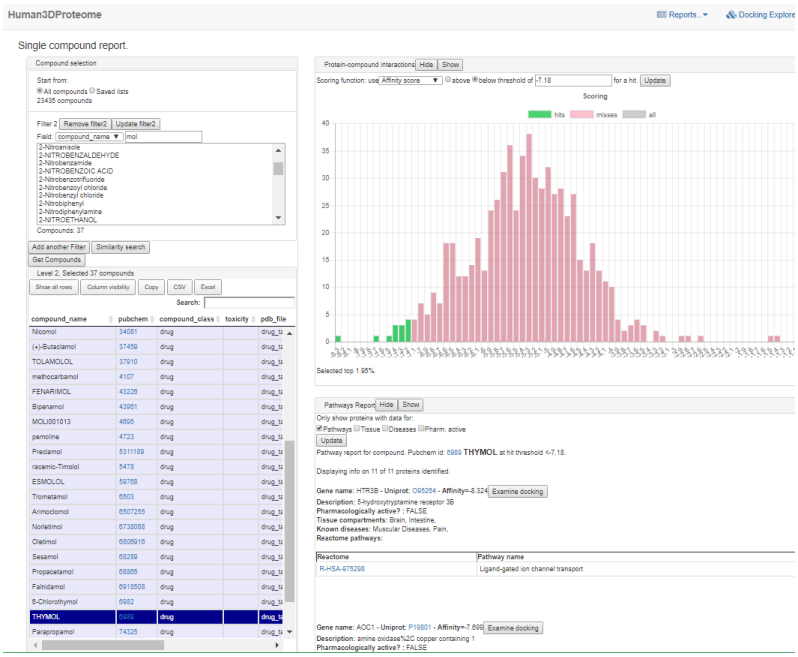
For dockings, a ‘pivot table’ style output is also given, allowing docking results to be visualised in the context of other dockings for that protein and compound. Column headings and data displayed can be changed, and table colouring schemes changed to visualise hits and misses, and help highlight results of interest.



Reports

Customised ‘report’ pages have been set up, to provide the user with a quick and simple way of exploring a single protein or compound. These use the same extensive docking database, with a user-definable threshold, to select ‘hits’ (the default thresholds can also be applied), and select biologically significant protein-ligand interactions. A graphical histogram shows the dockings selected as ‘hits’ out of all those in the database. The system also links proteins to the wider biological information known about them, such as tissue compartments and links to diseases, and their place in pathways in the Reactome database.

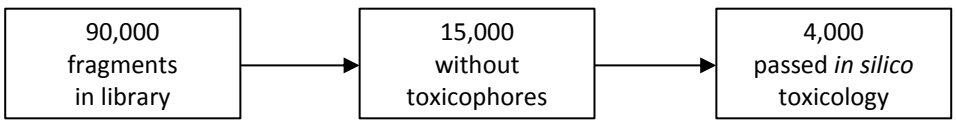
For example, a compound of interest may be linked, through dockings, to specific proteins, and from there to the pathways that might be affected, and the relevant therapeutic areas and diseases.



Workflows and case studies

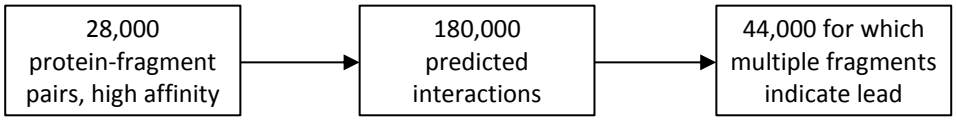
Demonstration of a Fragment-to-Lead Study

As a demonstration of the power of our pipeline, an example of a potential hit-to-lead approach was run. We use a fragment library pre-filtered for PAINS then further filtered by our *in silico* toxicity screen.



Of course, the wider set of 15,000 fragments (or even all 90,000 fragments) could be used for a more comprehensive search of chemical space if this is desired.

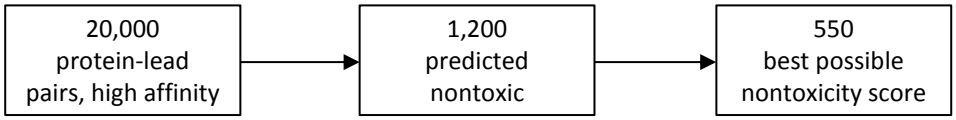
A panel of 64 common drug targets was assembled. Docking the 4,000 fragments to this panel resulted in 28,000 protein-fragment interactions of high affinity. By running a ligand similarity search, we identified 180,000 compounds from our ‘clean leads’ set that we assumed would potentially identify some interactions of similarly high affinity.



We limited this demonstration to investigating the 44,000 predicted interactions for which multiple protein-fragment pairs were indicated for the same protein-compound pair. The complete set of 180,000 predicted protein-compound interactions could be used if needed.

After molecular docking of the 44,000 protein-lead pairs, it was found that 20,000 of them had high affinity. All the compounds indicated as leads were screened through our *in silico* toxicology process.

The demonstration output was limited to taking only the 550 high-affinity protein-lead pairs for which the lead had the best possible non-toxicity score.



For validation purposes, we confirmed that 50 of these protein-lead interaction pairs involved a compound lead which was structurally most similar to the known FDA drug for that protein compared to all the other known drugs considered. These 50 drug-similar compound leads have therefore been independently identified from the fragment-to-lead approach comprising fragment docking, compound similarity searching, compound docking and toxicity screening. This is a good illustration of the power of the Human3DProteome platform applied to a specific question and workflow.

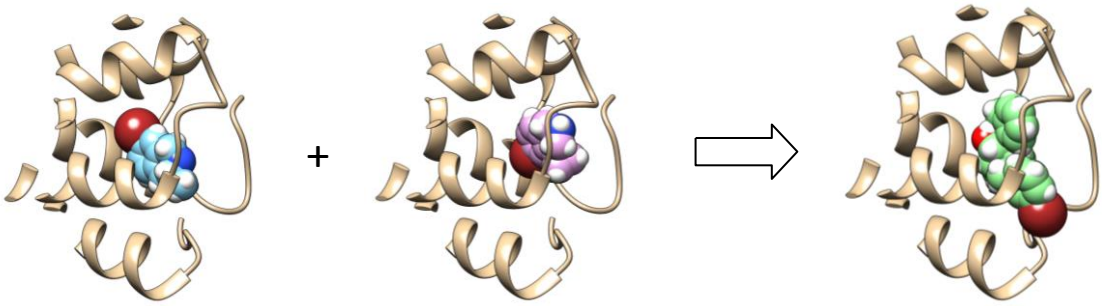
This data is presented in more depth per-protein in the following table:

Gene	Fragment Hits	Lead-Like Hits	Non-Toxic	Best Non-Toxic*	Similar to FDA
5HT1A	781	659	5	3	0
5HT1B	442	312	4	0	0
5HT1D	920	558	12	5	1
5HT2A	202	122	2	0	0
5HT2C	199	108	1	1	0
5HT6R	222	67	0	0	0
5HT7R	338	52	5	2	0
ABL1	967	377	3	0	0
ACM1	366	89	1	0	0
ACM2	635	416	5	2	1
ACM3	286	118	1	0	0
ACM5	359	290	5	1	0
ADA1B	412	162	10	4	0
ADA2B	328	136	0	0	0
ADA2C	199	70	1	0	0
ADRB2	114	41	0	0	0
AHR	933	954	79	34	2
AMPN	237	131	3	1	0
CA2D1	366	180	2	1	0
CA2D2	659	457	15	6	0
CADH5	510	373	26	8	0
CRBN	336	66	0	0	0
CSF1R	453	212	2	1	0
DDR1	860	886	41	14	2
DPP4	117	90	2	1	0

DRD1	623	199	1	0	0
DRD2	335	292	6	1	0
DRD4	286	101	2	1	1
DRD5	935	913	40	22	1
NMD3A	348	146	2	1	0
NMD3B	477	106	2	0	0
NMDE3	194	49	3	0	0
NMDE4	1598	1293	38	17	1
NPCL1	1054	544	16	6	2
NTRK1	384	310	17	3	0
PDE11	627	343	15	3	0
PDE5A	814	339	5	2	0
PGFRA	199	158	6	2	0
PGH2	1180	1060	29	13	0
SOAT1	322	144	1	0	0
SV2A	848	907	50	15	0
TNF11	1883	2765	533	318	42

This table has been filtered to show only the 42 receptors for which at least one lead-like hit was found.
* Best Non-Toxic indicates the lead-like compound achieved the safest possible toxicity rating.

Finally, we present an example output of the workflow. Here, two fragment dockings in ACM2 indicate the same lead.



The lead turns out to be structurally similar to Solifenacin, a muscarinic acetylcholine receptor antagonist.

A possible next step for identified leads is a whole-proteome screening to identify potential off-target interactions.

The platform is highly flexible and can be applied to multifarious drug discovery questions and approaches. At its heart is the powerful integration of all the molecular modelling and docking data with extensive biological and pharmacological data, all within a clearly structured, highly intuitive interface.

Human3DProteome provides structural modelling, proteome scale lead discovery, pharmacological profiling, toxicity screening, biological networks, *in silico* pathway analysis, graphical analysis and reporting, all in one place.

Visit Human3DProteome.com

Bibliography

Baell J.B. and Holloway G.A. (2010), "New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays," *J. Med. Chem.* **53**(7):2719-40.

Bowes J., *et al.* (2012), "Reducing safety-related drug attrition: the use of in vitro pharmacological profiling," *Nat. Rev. Drug Discov.* **11**(12):909-22.

Eurofins Cerep-Panlabs (2014), "SafetyScreen44," *http://www.cerep.fr/cerep/users/pages/downloads/Documents/Marketing/Pharmacology%20&%20ADME/Standard%20profiles/SafetyScreen44_2014v2LD.pdf*.

Fabregat A., *et al.* (2016), "The Reactome pathway Knowledgebase," *Nucleic Acids Res.* **44**(D1):D481-7.

Kim S., *et al.* (2016), "PubChem Substance and Compound databases," *Nucleic Acids Res.* **44**(D1):D1202-13.

Law V., *et al.* (2014), "DrugBank 4.0: shedding new light on drug metabolism," *Nucleic Acids Res.* **42**(D1):D1091-7.

Sipes N.S., *et al.* (2013), "Profiling 976 ToxCast chemicals across 331 enzymatic and receptor signaling assays," Chem. Res. Toxicol. **26**(6):878-95.

Sterling T. and Irwin J.J. (2015), "ZINC 15 – Ligand discovery for everyone," *J. Chem. Inf. Model.* **55**(11):2324-37.

Trott O. and Olson A.J. (2010), "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading," *J. Comput. Chem.* **31**(2):455-61.

Further reading

Please see http://scholar.google.co.uk/citations?user=e5sOfe0AAAAJ&hl=en for a catalogue of our academic research in this field.