To study interactions between monoclonal antibody and CHO Host Cell Proteins to design the wash process for mAb purification

A Major project submitted in fulfilment of requirement for the degree of

MASTER OF TECHNOLOGY

IN

BIOINFORMATICS

Submitted by

DIVYANSHI YADAV

(2K15/BIO/04)

Under the supervision of

Dr. Asmita Das
Department of Biotechnology
Delhi Technological University
Delhi-110042, INDIA.

CERTIFICATE

This is to certify that the M. Tech dissertation entitled **"To study interactions between monoclonal antibody and CHO Host Cell Proteins to design the wash process for mAb purification",** submitted by **DIVYANSHI YADAV (2K15/BIO/04)** in fulfilment of the requirements for the award of **Master of Technology in BIOINFORMATICS** and submitted to the Department of Biotechnology of Delhi Technological University (Formerly Delhi College of Engineering) is an authentic record of work carried out under the supervision of Dr. Asmita Das, Department of Biotechnology.

The information and data presented in this dissertation has not been submitted for the award of any other degree elsewhere.

.

DIVYANSHI YADAV
(Roll No. 2K15/BIO/04)

Dr. Asmita Das (Project Mentor)
Assistant Professor
Department of Biotechnology
Delhi Technological University

Prof. D. Kumar
Head of Department
Department of Biotechnology
Delhi Technological University

ACKNOWLEDGEMENT

It is a fact that every mission needs a spirit of hard work & dedication but it also needs to be put on right path. The credit goes to many well-wishers who have helped me to complete my dissertation work successfully with profound appreciation.

I would like to acknowledge my deep sense of gratitude to **Prof. D. Kumar,** Head of Department of Biotechnology, Delhi Technological University for giving me opportunity to study and work in this prestigious institute.

I owe my inexhaustible gratitude to **Dr. Asmita Das,** my mentor, Delhi Technological University who gave me a chance of doing this project work under the supervision of **Mr. Anurag Rathod**, IIT Delhi. I am very Thankful to **Mr. Sumit Kumar** for guiding me for the entire project and explaining so well. I will also like to thank all my faculty members and staff for their encouragement which helped me achieve success in my work.

I owe my sincere thanks to **Ayushi Garg, Shiva Rao and Mohini Yadav** who helped me throughout the course of the project work and helped me complete it in a timely manner.

DIVYANSHI YADAV
(Roll No. 2K15/BIO/04)

CONTENTS

ABBREVIATIONS USED

1) HCP	Host Cell Protein
2) mAb	Monoclonal Antibody
3) CHO	Chinese Hamster Ovary
4) PIR	Protein Information Resource
5) I-TASSER	Iterative Threading Assembly Refinement
6) SAVES	Structure Analysis and Verification Server
7) ProtSAv	Protein Structure Analysis and Validation
8) EMBL	European Molecular Biology Laboratory
9) EBI	European Bioinformatics Institute
10) NCBI	National Centre for Biotechnology Information

ABSTRACT

Today, recombinant protein therapies represent a substantial focus of the pharmaceutical industry. Most therapeutic proteins are produced in host cells of non-human origin, including Escherichia coli, yeast, and various mammalian cell lines e.g. Chinese Hamster Ovary Cell Lines. A major focus of all therapeutic protein purification process is to reduce components of host organism, including host cell proteins (HCPs), to levels considered as adequate in the final formulated drug product. HCPs can pose potential safety risks for patients, including immune reactions, decreased product stability, adjuvant activity, and (theoretically) protein-specific biological activity.(Schenauer, Flynn, & Goetze, 2012)

With rapidly growing cases and increased number of cancer, autoimmune diseases, Alzheimer's disease has become the most common death-causing diseases worldwide. Recent studies indicate that mAb is effective in treatment of these diseases and with limited number of treatment options people need to rely on these medicines. Thus, the purity of these medicines becomes an important factor. If these medicines are not pure the HCP might induce antigenic reactions in the patient also these HCPs if proteolytic may degrade the desired amount of mAb to be effective as dose, thus making the medicine ineffective over a period. Thus here, we report results of the studies relating to the most interactive HCPs which isolated along with mAb and studied their interactions to design the wash process accordingly. Using the proteome of Chinese Hamster, the hotspot for the proteins were found using Aggrescan. Structure prediction was done using threading software's Bhageerath, Pyre 2, Multicom-Raptor-X, Robetta and I-TASSER. The model generated was further validated by Independent servers Prochek, Verify-3d, Errat and with meta servers such as Protsav and SAVES. The models were refined using 3D refine and galaxy refine.

The best models were docked with mAb using cluspro to identify the interactions between HCP and mAb then find the amino acid interaction profile using PDB-Sum and further work on developing modified wash process to break these bonds and obtain pure mAb to ensure effective treatment by functionally characterizing the proteins.

INTRODUCTION

Monoclonal Antibodies are produced by cell division from a single ancestral cell. These are relatively specific for a location in the body and they can be grown indeterminately. Monoclonal Antibodies recognize and bind to antigens to distinguish between specific epitopes which provide defence against disease organisms.

Monoclonal antibodies target several proteins that influence cell activity such as some proteins or receptors present on the surface of cancerous and normal cells. The specificity of mAbs allows it to bind to cancerous cells by coupling a cytotoxic agent such as a radioactive which then seek outs to destroy the cancer cells without harming the healthy cells.

Tumor cells that can replicate endlessly are fused with mammalian cells so that they produce a specific antibody which result in the fusion called hybridoma that continuously produce antibodies. These antibodies are termed monoclonal because they come from only single type of cell, which is a hybridoma cell.

Monoclonal antibodies are artificially produced against a specific antigen to bind to their target antigens.

Monoclonal Antibodies are much more effective than conventional drugs since the drugs attack the foreign substance & also the body's own cells thus, causes harsh side effects but the monoclonal antibodies only target the foreign antigen/target molecule, without or with only some minor side effects.

Production of Monoclonal Antibodies:

Large quantities of targeted antibodies against a specific antigen are produced via multiple identical copies of a certain cell called hybridoma. For creating Hybridoma cells fusion of two cells is needed to combine the characteristics of the two cells into one cell. One of them produces antibody cells which is a B-Lymphocyte from a laboratory hamster and the other is a tumor cell termed myeloma.

Tumor cells can grow indefinitely and at an exceeding rate from normal cell growth. Laboratory produced Hybridoma cells replicate much faster than normal antibody producing cells, and the hybridomas produce the specific antibodies for an indefinite period.

All monoclonal antibodies have 'mab' at the end of their generic name, few commonly used mAb in treatments are for example: trastuzumab (Herceptin), bevacizumab (Avastin), Rituximab etc.

The Chinese hamster (***Cricetulus griseus syn. Cricetulus barabensis griseus***) are a species of Cricetidae they originate from the deserts of Mongolia and Northern China. The entire 16,284

base pair nucleotide sequence of the Chinese hamster mitochondrial genome was published in 2007. In 2011, it was followed by the genome sequencing and annotation of the ancestral CHO-K1 cell lines. Analysis of the glycosylation genes in the CHO-K1 genome found 99% homologs of the human glycosylation-associated transcripts, 53% of them are expressed. Chinese Hamster Ovary cells (CHO) cells are the predominant host used to produce therapeutic proteins. About 70% of all recombinant proteins today are produced in CHO cells. Their ability to grow to high density in serum-free suspension culture that are readily scaled to >10,000-L bioreactors, as well as to secrete and express proteins with the appropriate human-compatible post-translational modifications (e.g., glycosylation). Also, that recombinant CHO cells can successfully grow in large-scale cultures of either suspension-adapted cells or adherent cells. Also, in a study in 1989 it was found that, out of 44 human pathogenic viruses tested, many them including HIV, polio, herpes, influenza, and measles do not replicate in CHO.(Li, Vijayasankaran, Shen, Kiss, & Amanullah, 2010)

CHO cells also have a proven track record for producing proteins with glycoforms that are both bioactive and compatible in humans. They have been demonstrated as a safe host for synthesis of biologics. Downstream processes for CHO cell products have mellowed to a stage where they can be purified to contain picogram levels of contaminating CHO Host Cell Proteins (HCPs) per dose of the product. (Wlaschin & Yap, 1987)

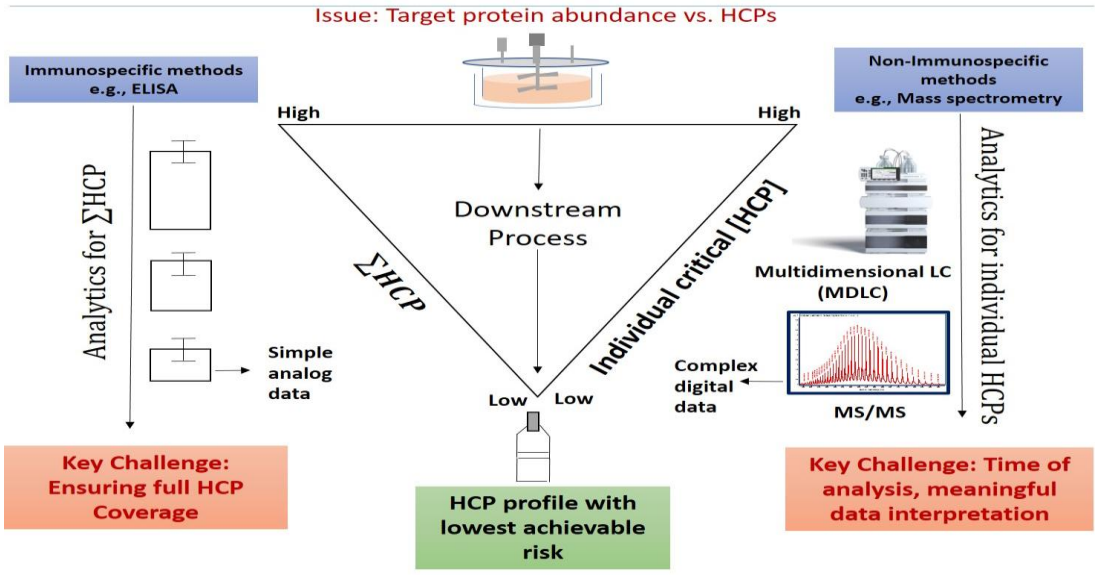


Figure 1: Different methods for identifying HCP's and problems related to it

There are various HCP's reported in the harvesting of the mAb from the CHO cell lines although washing is done but it still HCPs are present in the harvest and this study is done to facilitate the removal of these HCPs.

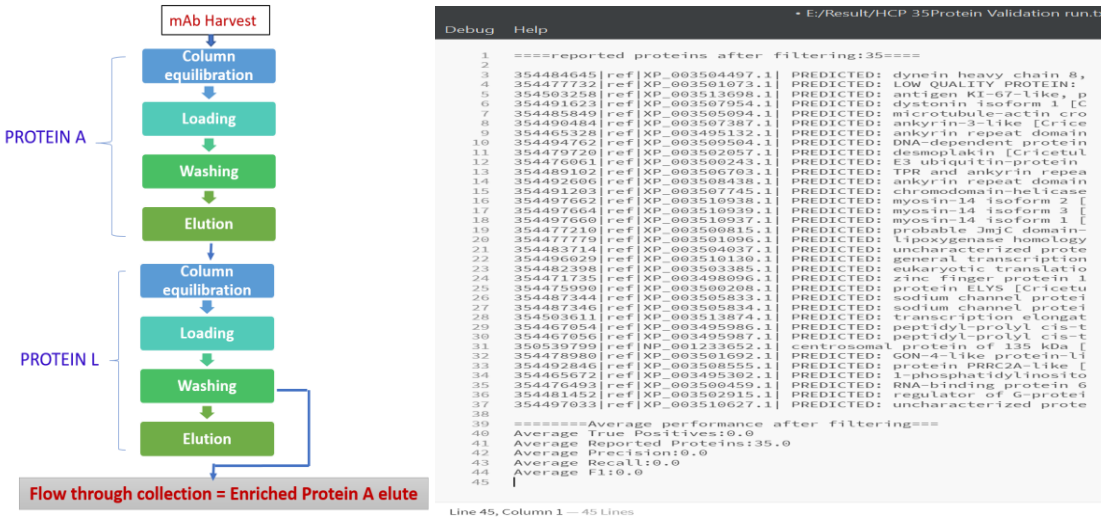


Figure 2: Steps involved in mAb purification and different HCP's being reported after washing and filtration

HCP'ˢ are a major problem in purification of mAb'ˢ.

1. They coelute with the mAb.
2. HCP's could be of varied nature but most harmful are immunogenic or proteases in nature as if they coelute with our mAb they will degrade it over time.
3. Thus, reducing its effective doses and ultimately making them ineffective for a treatment.
4. Immunogenic HCP'ˢ cause immune response in the patient making their condition worse.

Thus, their further study and removal of the existing HCP's is needed by modifying the existing wash processes which could be done by studying the kind of interactions occurring between the HCP's and the Mab.

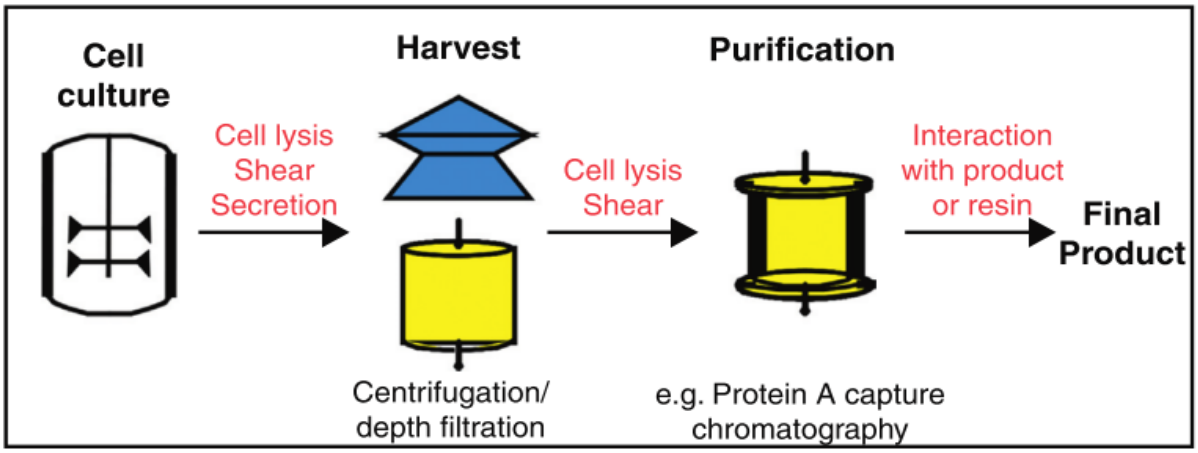


Figure 3: Different steps involved in Purification and isolation of mAb

REVIEW OF LITERATURE

The immune system in vertebrates is endlessly evolving to protect itself from diverse intruding pathogens. The immune responses revolve around some innate mechanisms, including adaptive processes such as producing antibodies (Ab) that can bind to all molecular structures of the microbial pathogen (bacteria, viruses, fungi, nematodes, and other parasites) and can keep pace with the diversified mutations in an organism. An antigen is defined as a molecule or part of a molecule that can be recognized by the immune system as a foreign entity. The challenge of the immune system is thus combated in two ways. First, through an antibody diversity mechanism, B lymphocytes produce varied antibodies specific for a new antigen (epitope) expressed by a pathogen by shuffling and reshuffling its genetic constituents. Second, paratope-encoding genes of the antibody are mutated rapidly to cope and bind strongly with the epitope of the antigen. Thus, these produced antibodies are better at binding with the antigen with greater affinity and high specificity.

Therefore, antibodies are useful research tools in diagnosis and therapy, as they can recognize and bind specifically and strongly with respective antigens.

Monoclonal antibodies (MAb(s)) are a mixture of homogenous antibody molecules with affinity towards a specific antigen, often generated using a hybridoma by fusing a B-cell with a single lineage of cells containing a definite antibody gene. Finally, a population of identical cells (or clones) is produced that secrete the same antibody.

Due to their specificity and high reproducibility using culture techniques, MAbs offer advantage over polyclonal antibodies. MAbs are increasingly used in applications such as research and diagnosis, therapeutic tools in cancer and immunological disorders, and pharmacy, thus generating a great demand in industry. The essential characteristics that confer the clinical applicability of MAbs include their specificity of binding and homogeneity, as well as their ability to be produced in unlimited quantities. Another unique advantage of hybridoma production is that a mixture of antigens can be employed to generate specific antibodies. This also enables one to screen an antibody of choice from a mixture of antibody population with a purified antigen; thus, a single cell clone can be isolated.

History

The production of MAbs by hybridoma technology was discovered in 1975 by Georges Kohler of West Germany and Cesar Milstein of Argentina, who jointly with Niels Kaj Jerne of Denmark were awarded the Nobel Prize for Physiology and Medicine in 1984. In 1976, Kohler and Milstein developed a technique to fuse splenocyte cells (separated from the spleen of an immunized mouse) with tumorous myeloma cells. The hybrid cells were clones

of antibody producing cells against a desired antigen and propagate rapidly to produce very large amounts of antibody. The hybridoma is capable of rapid propagation and high antibody secreting rates such as in myeloma cells, which can maintain the antibody genes of mouse spleen cells. In 1988, Greg Winter used the first humanized MAbs to avoid reactions/responses observed in patients injected with murine derived MAbs.1–4(Ansar & Ghosh, n.d.)

Outline of production of MAbs

The main objective is to produce a homogenous population of MAbs against a pre-fixed immunogen. The basic strategy includes:

(i) Purification and characterization of the desired expression vector,

(ii) Choosing the CHO cell lines.

(iii) Transfecting the expression vector into the cell line

(iv) After transfection select the hybrid cell line with gene of interest by selection marker at increased concentration.

(v) Clone the desired cell to get bulk quantity.

(vi) The antibodies secreted by the different clones are then tested for their ability to bind to the antigen using an enzyme-linked immunosorbent assay (ELISA).

(vii) The clone is then selected for future use.

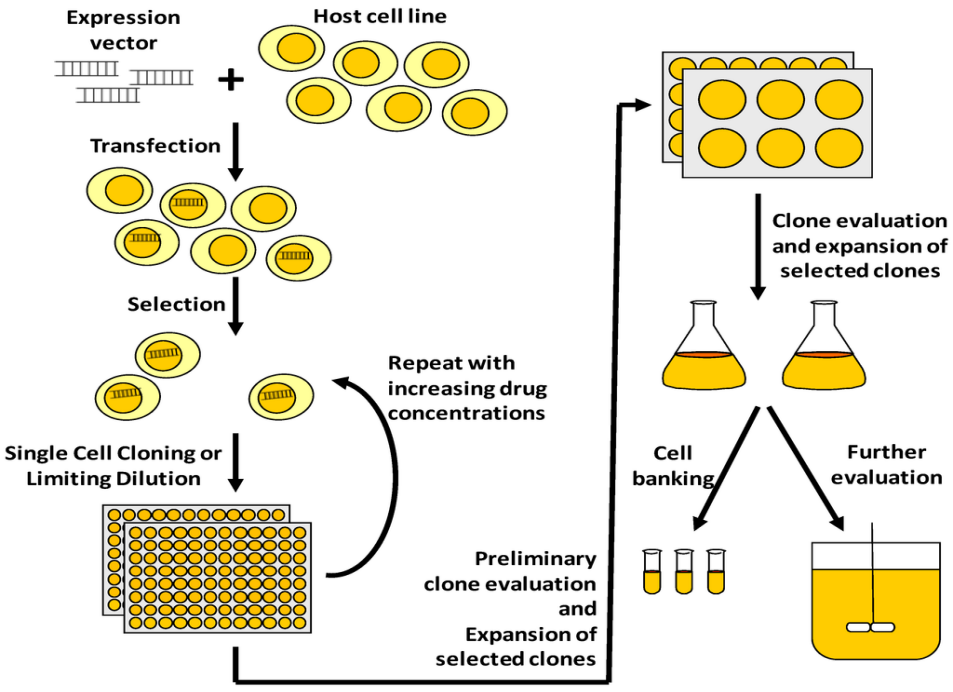


Figure 4: Different steps for the creation of Chinese Hamster Ovary Cell lines to produce mAbs

Therapeutic proteins, produced in genetically modified host cells, represent a growing class of treatments within the biopharmaceutical industry, Chinese hamster ovary (CHO) cells are the primary expression system for therapeutic proteins, which are typically secreted into the extracellular medium along with endogenous host cell protein (HCP) impurities that must be removed from the product for patient safety.

Identification and characterization of these extracellular CHO HCPs by proteomic techniques can aid bioprocess development, resulting in robust biopharmaceutical manufacturing operations. Developing optimized sample preparation protocols that improve extracellular CHO HCP capture is fundamental for maximizing the utility of these proteomic methods. (Valente, Schaefer, Kempton, Lenhoff, & Lee, 2014)

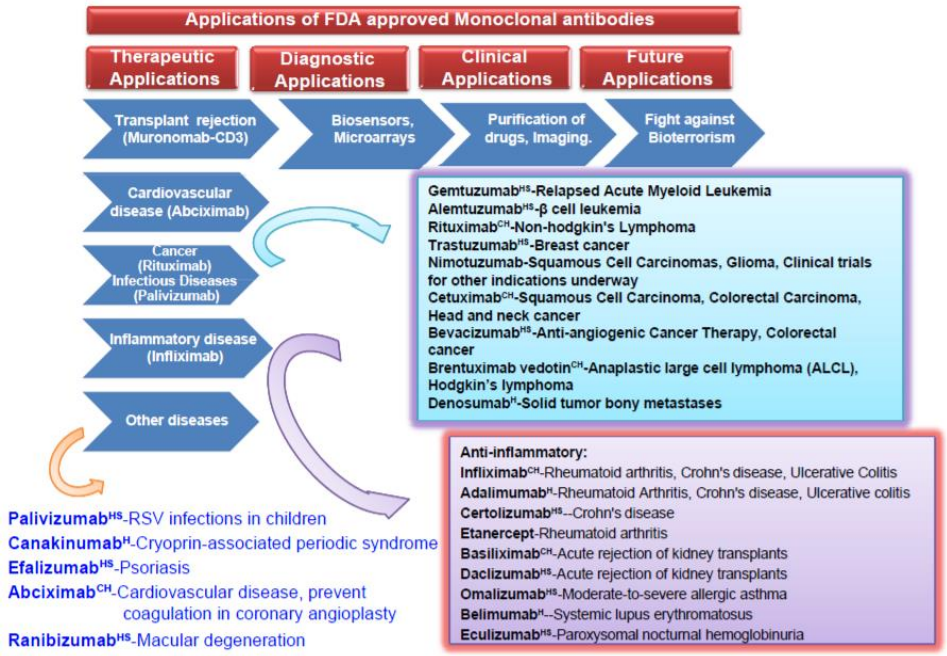


Figure 5: Applications of certain FDA approved mAbs

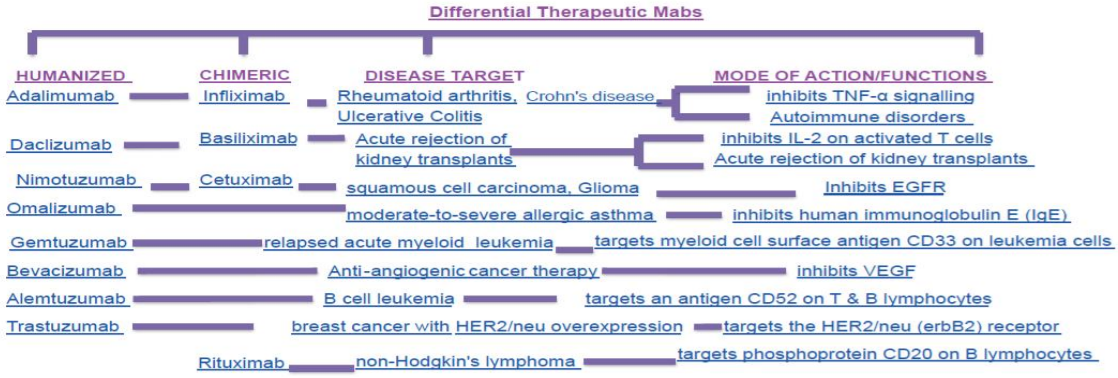


Figure 6: Mode of action of different therapeutic mAbs

Chinese Hamster Ovary cells (CHO) are commonly used for expression of recombinant proteins, including monoclonal antibodies.

Production of recombinant therapeutic proteins is inextricably associated with co expression of endogenous proteins by the host cell, so called host cell proteins (HCPs). Made up by a multiplicity of proteins that possess a broad variety of physicochemical and immunological features, HCPs represent a major process-related impurity. Most products evaluated and approved by health authorities contains HCPs at levels lower than 100 parts per million (ppm), since HCP may potentially trigger adverse effects such as eliciting immune responses against the HCP itself or against the therapeutic drug product. Assessment of HCP contaminants in protein drug solutions requires a highly sensitive and specific analytical test method. The predominant test method fulfilling this requirement is typically an ELISA, which represents the well-established standard for the quantification of HCP content in all bioprocess production steps. The favourable characteristics of this immunoassay format include the sensitivity, precision and cost–effectiveness required to support valid determination of process-related impurity levels. However, this method suffers from drawbacks such as total procedure time and operator workload. Contemporary, Quality by Design (QbD)-driven process development approaches and process characterization and validation studies result in high sample numbers. (Leiss, Pester, & Aschner, 2015)

TOOLS AND DATABASES USED

1) Aggrescan

AGGRESCAN is an easy to use web-server that permits the simultaneous analysis of the aggregation properties of large number of proteins in very less time, independent of their size. It also provides graphs to facilitate easy identification of the distribution of likely to aggregate residues in a polypeptide sequence. Aggrescan can be widely used to identify antibodies that are able to block protein aggregation in disease related processes, regions that are able to interact with excipients of therapeutically relevant proteins during storage and thus increase their shelf life, find putative substrates for molecular chaperones. To find information about the cytotoxic mechanism of a proteins, To improve the solubility of therapeutic proteins etc.(Conchillo-Solé et al., 2007)

2) Verify 3D

Verify 3D determines the compatibility of an atomic three-dimensional models with its own amino acid sequence (1-Dimentional) by assigning a structural class based on its environment and location (beta, alpha, loop, nonpolar, polar, etc) and comparing the results to good structures. (Bowie, et al., 1991; Luethy, et al., 1992)

3) Phyre 2

Phyre2 is a suite of different tools that are available on the web server (http://www.sbg.bio.ic.ac.uk/phyre2) to analyse and predict protein function, structure, and mutations. Phyre2, uses advanced remote homology detection methods to build 3D models, and to analyse the effect of amino acid variants and predict ligand binding sites for a given protein sequence. Results are displayed by a simple interface with details a user can determine.(Kelley, Mezulis, Yates, Wass, & Sternberg, 2015)

4) UniProt

The Universal Protein Resource (UniProt) is a wide-ranging resource for protein sequence and annotation data. The UniProt databases are the UniProt Knowledgebase (UniProtKB), the UniProt Archive (UniParc) and the UniProt Reference Clusters (UniRef). UniProt is a collaboration between the SIB Swiss Institute of Bioinformatics, the European Bioinformatics Institute (EMBL-EBI), and the Protein Information Resource (PIR). SIB and EMBL-EBI together produce TrEMBL and Swiss-Prot, while PIR produce the Protein Sequence Database (PIR-PSD). These two data sets coexisted with annotation priorities and different protein sequence coverage. A proteome is the set of proteins expressed by an organism. UniProt provides proteomes of species with wholly sequenced genomes. Some proteomes have been (algorithmically and manually) chosen as reference proteomes. They cover well-studied prototype organisms and other organisms of interest for biomedical research and phylogeny.

5) Bhageerath

Bhageerath-H is a software suite for protein tertiary structure prediction using their amino acid sequence in FASTA format and predicts five probable near native structures. The software encompasses seven computational modules which work in channel and together form an automated pipeline. Following sections discuss each module of the automated pipeline. The very first step in the pipeline involves generation of a bulk pool of full length decoys, followed by BLAST and secondary structure prediction. It exploits amino acid chemical properties such as conformational flexibility, hydrogen bond donor, size and shape of side chains for generating an amino acid substitution scoring matrix. This scoring matrix is used for template-target alignment generation as well as template selection. This matrix helps in selecting distant homologs, which are usually missed in a normal database search. The template-target and template alignments are used for modeling fragments of varying length via Modeller. These, modelled fragments are then screened for missing links. These missing stretches are created using Bhageerath ab initio modeling method. Then a single representative structure of each unique topology is retained. Then this set of decoys containing near-native models is submitted for physico-chemical scoring in the third step. To get the top 5 structures, each of these seven parameters are evaluated and the structure is predicted. (Jayaram et al., 2014)

6) Robetta

The Robetta server (http://robetta.bakerlab.org) provides an automated tool for protein structure prediction and analysis. For structure prediction, amino acid sequence is

submitted to the server and are parsed into putative domains and the structural models are generated using either de novo or comparative modeling structure prediction methods. Robetta server uses the first completely automated structure prediction method that produces a model for a complete protein sequence in the absence or presence of sequence homology to protein(s) of already known structure. Robetta builds models for domains with sequence homology of known protein structures using comparative modeling, and models for domains missing such homology using the Robetta de novo structure prediction method.(Kim, Chivian, & Baker, 2004)

7) Multicom Raptor-X

RaptorX server (http://raptorx.uchicago.edu) is for protein secondary structure prediction, it uses template-based tertiary structure modeling, and probabilistic alignment sampling. RaptorX server can detect remotely related template sequence for a given sequence by a novel nonlinear context-specific probabilistic consistency algorithm and alignment potential. It thus makes it possible to obtain high-quality structural models for numerous target protein sequences where only distantly related protein domains are having experimentally solved structures. The predicted 3D models can be used in protein docking and protein–protein interaction studies as well as for binding site epitope prediction.(Cheng, Li, Wang, Eickholt, & Deng, 2012)

8) I Tasser

I-TASSER server (http://zhang.bioinformatics.ku.edu/I-TASSER) is a unified platform for automated protein structure prediction and function prediction based on their sequence-to-structure-to-function paradigm. I-TASSER first generates 3-d atomic models from iterative structural assembly simulations and multiple threading alignments. The protein function is then inferred by structurally matching the three-dimensional models with other known protein structures. The output contains full-length tertiary and secondary structure predictions, and functional annotations on ligand-binding sites, It's methodology is divided into four general stages. The first stage is threading which refers to a procedure for identifying template proteins from experimentally verified structure databases. Later the templates are ranked by a variety of structure-based and sequence-based scores. The best template from each threading program are then selected intended for further consideration. In the second stage, structural assembly is done by excising continuous fragments in threading alignments from the template structures, and then are used to assemble structural conformations aligned well, with the unaligned regions built by ab-initio modeling. In the third stage, the fragment assembly simulation is performed again right from the selected cluster centroids. The purpose of this second iteration is to refine the global topology of the cluster centroids as well as to remove steric clashes. The structures generated during the second round of simulations are again clustered, and lowest energy structures are selected to generate the final structural models by building all-atom models through the optimization of hydrogen-bonding networks. In final stage, the function of the query protein by matching the predicted 3D models with the proteins of known function and structure in the PDB.(Yang et al., 2015)

9) Procheck

The PROCHECK suite provides a detailed check about the stereochemistry of protein structures. The output comprises of a comprehensive residue-by-residue listing and several plots in PostScript format. These give a evaluation of an overall quality of structure as compared to the refined structures of the same resolution and also highlighting the regions that may need further study.

The PROCHECK suite comprises of five programs, the PROCHECK suite is easy to use and proved useful for the solutions of new structures, model building of unknown structures and assessment of existing structures. (Thornton, 1993)

10) Errat

ERRAT is a program for which verifies protein structures determined by crystallography. The error function depends on the statistics of non-bonded atom-atom interactions in the reported structure (comparing to a database of reliable high-resolution structures). The figure shows a plot of the initial and final model. Regions of the structure that could be rejected at the 95% confidence level are yellow in colour; 5% of the good protein structure is expected to be having an error value above this level. Regions that could be rejected at the 99% level are shown in red colour. According to the ERRAT analysis, the final model is significantly improved comparative to the initial model.(Colovos & Yeates, 1993)

11) PROTSAV

ProTSAV is open access (http://www.scfbio-iitd.res.in/software/proteomics/protsav.jsp). It is, capable of evaluating the predicted model structures based on some popular online standalone tools and servers. It equips the user with a single quality score in case of an individual protein structure along with the graphical representation and ranking in case of multiple protein structure valuation. The server is validated on approximately 64,446 protein structures including predicted model structures for CASP targets and experimental structures from RCSB and from public decoy sets. ProTSAV succeeded in predicting quality of protein structures with a 100% specificity and 98% sensitivity on experimentally solved structures and achieves 88% specificity and 91% sensitivity on predicted protein structures of CASP11 targets below 2 Å. (Singh, Kaushik, Mishra, Shanker, & Jayaram, 2016)

12) Cluspro

The ClusPro server (https://cluspro.org) is an extensively used tool for protein–protein docking. The server requires only two files in Protein Data Bank (PDB) format, or by providing the PDB Id's. ClusPro offers a number of advanced options to modify the search; such as removal of unstructured protein regions, accounting for pairwise distance restraints, application of attraction or repulsion, location of heparin-binding sites and construction of homo-multimers, consideration of small-angle X-ray scattering (SASAXS) data. Depending on the type of protein six different energy functions can be used. (Kozakov et al., 2017)

13) SAVES

This meta-server runs six programs for validating and checking protein structures during and after model refinement.

a) **PROCHECK**: It checks the stereochemical quality of a protein structure by analysing overall and residue-by-residue geometry.

b) **WHAT_CHECK**: It is derived from a subset of protein verification tool WHATIF program (Vriend, 1990), this does extensive checking of the residues in the model for many stereochemical parameters.

c) **ERRAT**: It analyzes the statistics of non-bonded interactions amongst different atom types and plots the value of error function versus position of a 9-residue sliding window, calculated by a comparing statistic from highly refined structures.

d) **VERIFY_3D**: It determines the compatibility of an atomic three-dimensional model with its own amino acid sequence (1-Dimentional) by assigning a structural class based on its environment and environment (beta, alpha, loop, nonpolar, polar, etc) and comparing the results to good structures. (Bowie, et al., 1991; Luethy, et al., 1992)

e) **PROVE**: It calculates the volume of atoms in macromolecules using an algorithm which treats the atoms like a hard sphere and calculate a statistical Z-score deviation for the models from highly refined (R-factor of 0.2 or better) and resolved (2.0 Å or better) PDB-deposited structures.

f) **CRYST1**: CRYST1 record and search the entire Protein Data Bank for the matches and report these as possibly similar structures.

g) **Ramachandran Plot**: The Ramachandran plot is used to evaluate structures and to find whether the main chain torsion angles phi-psi (φ,ψ) torsion angles for all residues in the structure (except those at the chain termini) are stereochemically feasible. The different regions on the Ramachandran plot are as described in Morris et al. (1992)

One can run all six programs to get a collective view of the input structure, or also individual programs can be selected.

14) PDB-SUM

The PDB-SUM server performs three computational steps

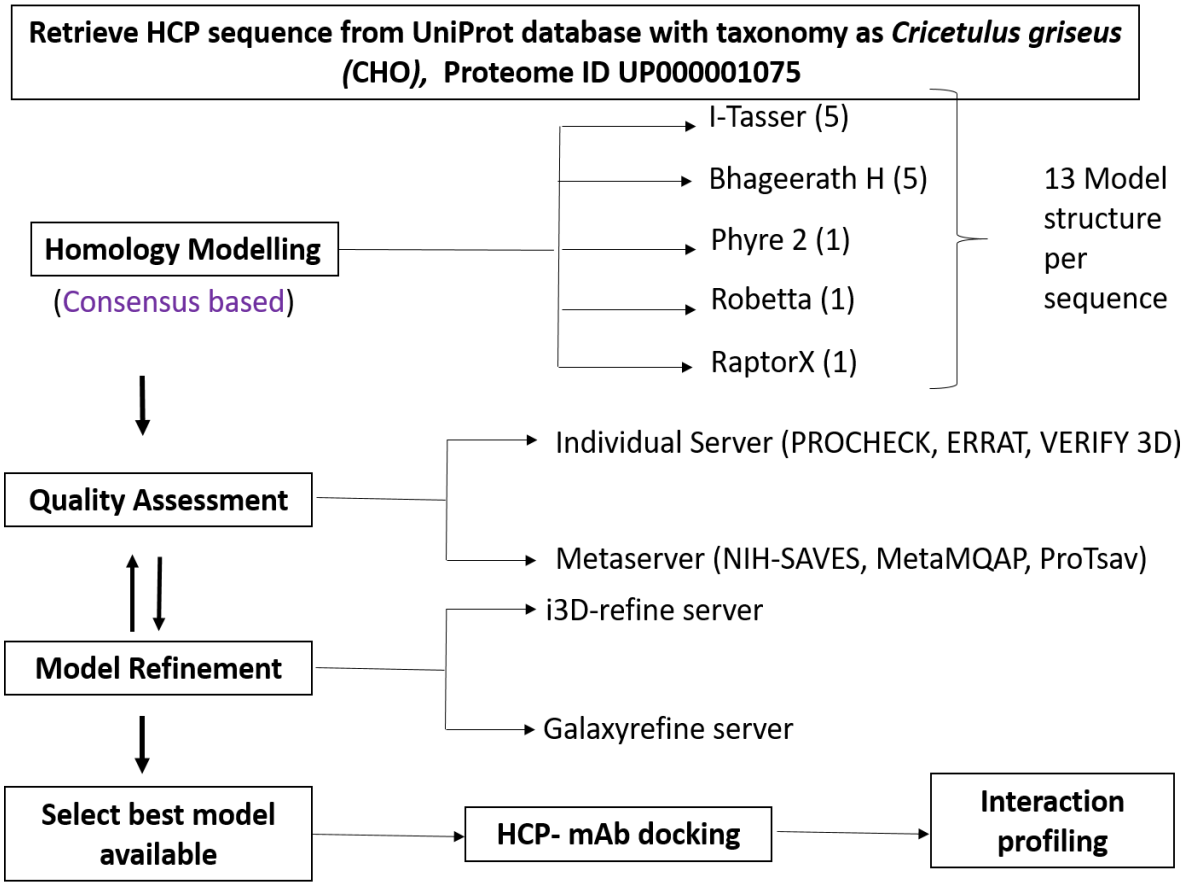
(a) Rigid-body docking: by sampling billions of conformations

(b) Refinement of selected structures using energy minimization

(c) Root-mean-square deviation (RMSD): based clustering of the thousand lowest-energy structures produced, to find the major clusters that will denote the most likely models of the complex.(Laskowski, 2001)

MATERIAL AND METHODS

WorkFlow



The complete workflow which was followed was divided into different subparts making it possible to do the complete analysis.

1) **Retrieval of CHO Proteome**

   The Chinese hamster ovary [*Cricetulus griseus*] whose protein structures are not yet available in RCSB Protein Databank (PDB) was retrieved from CHO Genome (http://www.chogenome.org/index.php) and referred with Uniprot Proteome Database (http://www.uniprot.org/proteomes/UP000001075) that houses a series of databases relevant to biotechnology and biomedicine in FASTA format.

2) **Literature search for the reported HCP's**

   HCP's eluting with the mAb during its purification are quite well studied in literature in different prospectives. Data collection was done by reading different research papers of the similar work and curating the repeated data from NCBI resource Pubmed (https://www.ncbi.nlm.nih.gov/pubmed/).

3) **Classification of the reported HCP's**

The experimentally verified HCP's were categorized based on their function into three parts.

i. **Product Associated HCP's:** HCP's isolated with the downstream processesing via direct interactions with the mAb.(Levy, Valente, Choe, Lee, & Lenhoff, 2014)

ii. **Varying Expression HCP's:** Proteins that co-purify due to their retention characteristics on chromatographic media through strongly attractive interactions to the therapeutic protein. Two complementary proteomic techniques can be used to identify HCPs with variable expression levels two-dimensional electrophoresis (2DE) and shotgun.(Valente, Lenhoff, & Lee, 2015)

iii. **Co eluting HCP's:** The HCP's that are left as residues during purification of mAb.

Product association and co-elution are both identified as most viable mechanisms of HCP retention. These three categories of HCP's were analysed together to identify the HCP's exhibiting all three characteristics thus they will be the most difficult to remove HCP's and should be studied further in order to identify the kind of interactions they are forming with the mAb's and accordingly the wash could be designed to remove these HCP's. (Levy, Valente, Lee, & Lenhoff, 2016)

4) **Hotspot Identification**

Hotspot study of the CHO Proteome was done using the tool Aggrescan (http://bioinf.uab.es/aggrescan/) to identify the number of hotspots present in each protein of the CHO for its further analysis.(Conchillo-Solé et al., 2007)

5) **Model Building**

The three-dimensional structure of protein was first modelled using the FASTA sequence with five different Web servers for Protein Tertiary Structure Prediction

i. **Bhageerath:** Bhageerath-H" accepts amino acid sequences in FASTA format to predict 5 candidate structures for the native. The amino acid sequence can be either pasted or typed in the box. The current version of Bhageerath supports upto 100 amino acid sequences. Bhageerath-H is a software suite for protein tertiary structure prediction using their amino acid sequence in FASTA format and predicts five probable near native structures. The software encompasses seven computational modules which work in channel and together form an automated pipeline. Following

sections discuss each module of the automated pipeline. The very first step in the pipeline involves generation of a bulk pool of full length decoys, followed by BLAST and secondary structure prediction. It exploits amino acid chemical properties such as conformational flexibility, hydrogen bond donor, size and shape of side chains for generating an amino acid substitution scoring matrix. This scoring matrix is used for template-target alignment generation as well as template selection. This matrix helps in selecting distant homologs, which are usually missed in a normal database search. The template-target and template alignments are used for modeling fragments of varying length via Modeller. These, modelled fragments are then screened for missing links. These missing stretches are created using Bhageerath ab initio modeling method. Then a single representative structure of each unique topology is retained. Then this set of decoys containing near-native models is submitted for physico-chemical scoring in the third step. To get the top 5 structures, each of these seven parameters are evaluated and the structure is predicted. (Jayaram et al., 2014)

ii. **Phyre 2:** Phyre2 is a suite of different tools that are available on the web server (http://www.sbg.bio.ic.ac.uk/phyre2) to analyse and predict protein function, structure, and mutations. Phyre2, uses advanced remote homology detection methods to build 3D models, and to analyse the effect of amino acid variants and predict ligand binding sites for a given protein sequence. Results are displayed by a simple interface with details a user can determine. (Kelley et al., 2015)

iii. **I-TASSER:** I-TASSER server (http://zhang.bioinformatics.ku.edu/I-TASSER) is a unified platform for automated protein structure prediction and function prediction based on their sequence-to-structure-to-function paradigm. I-TASSER first generates 3-d atomic models from iterative structural assembly simulations and multiple threading alignments. The protein function is then inferred by structurally matching the three-dimensional models with other known protein structures. The output contains full-length tertiary and secondary structure predictions, and functional annotations on ligand-binding sites, It's methodology is divided into four general stages. The first stage is threading which refers to a procedure for identifying template proteins from experimentally verified structure databases. Later the templates are ranked by a variety of structure-based and sequence-based scores. The best template from each threading program are then selected intended for further consideration. In

the second stage, structural assembly is done by excising continuous fragments in threading alignments from the template structures, and then are used to assemble structural conformations aligned well, with the unaligned regions built by ab-initio modeling. In the third stage, the fragment assembly simulation is performed again right from the selected cluster centroids. The purpose of this second iteration is to refine the global topology of the cluster centroids as well as to remove steric clashes. The structures generated during the second round of simulations are again clustered, and lowest energy structures are selected to generate the final structural models by building all-atom models through the optimization of hydrogen-bonding networks. In final stage, the function of the query protein by matching the predicted 3D models with the proteins of known function and structure in the PDB.(Yang et al., 2015)

iv. **Robetta:** The Robetta server (http://robetta.bakerlab.org) provides an automated tool for protein structure prediction and analysis. For structure prediction, amino acid sequence is submitted to the server and are parsed into putative domains and the structural models are generated using either de novo or comparative modeling structure prediction methods. Robetta server uses the first completely automated structure prediction method that produces a model for a complete protein sequence in the absence or presence of sequence homology to protein(s) of already known structure. Robetta builds models for domains with sequence homology of known protein structures using comparative modeling, and models for domains missing such homology using the Robetta de novo structure prediction method.(Kim et al., 2004)

v. **Multicom Raptor-X:** RaptorX server (http://raptorx.uchicago.edu) is for protein secondary structure prediction, it uses template-based tertiary structure modeling, and probabilistic alignment sampling. RaptorX server can detect remotely related template sequence for a given sequence by a novel nonlinear context-specific probabilistic consistency algorithm and alignment potential. It thus makes it possible to obtain high-quality structural models for numerous target protein sequences where only distantly related protein domains are having experimentally solved structures. The predicted 3D models can be used in protein docking and protein–protein interaction studies as well as for binding site epitope prediction.(Cheng et al., 2012)

6) **Structure Validation**

To validate model structures individual servers and meta servers were used. Ramachandran Plot of the best model was obtained using UCLA MBI. Structure Analysis and Verification Server (SAVES) (http://services.mbi.ucla.edu/SAVES/Ramachandran/). Further Stereochemical validation was done using UCLA MBI PROCHECK (http://services.mbi.ucla.edu/PROCHECK/). It gives an idea about overall quality of the structure by comparing with well refined structures of the same resolution and highlight regions that may need further investigation. UCLA MBI ERRAT (http://services.mbi.ucla.edu/ERRAT/) was used for analyzing the statistics of non-bonded interactions between different atom types. UCLA MBI Verify3D (http://services.mbi.ucla.edu/Verify_3D/) was also used for determining the compatibility of the generated model (3-Dimentional) with its own amino acid sequence (loop, polar, 1-Dimentional) by assigning a structural class based on its environment and location (alpha, beta, nonpolar etc.) and comparing the results to good structures.

Another bioinformatics tool Protein Structure Analysis and Validation (ProTSAV) (http://www.scfbio-iitd.res.in/software/proteomics/protsav.jsp) was used which is a meta-server, and has a collection of model quality assessment programs that evaluate the correctness of the structural model and quality of a protein. It also predicts a global quality score derived from quality assessment of different validation tools (modules) selected by user.

7) **Energy Minimization**

Energy minimization of models was done using galaxy refine (http://galaxy.seoklab.org/cgi-bin/submit.cgi?type=REFINE) and 3-d refine (http://sysbio.rnet.missouri.edu/3Drefine/) to obtain most stable energy conformation of the protein molecule.

8) **Docking:** Protein–protein docking simulations were performed using Cluspro protein–protein docking web server v. 2.0 (Kozakov et al., 2010). The program recruits PIPER, a Fast Fourier Transform (FFT) based rigid docking program in its initial stage to generate 1,000 low energy docked conformations using pairwise interaction potentials. In its second stage, ClusPro cluster these conformations and retains 30 largest clusters having lowest energy. Later, the retained clusters are analyzed by Semi-Definite programming based Underestimation (SDU) program, it predicts the stability of the clusters using

medium-range optimization algorithm (resembles to funnel-like behaviour of free energy to attain local minima) and these stable clusters are then refined using Monte-Carlo simulation. The server performs mainly three computational steps as follows:

(i) rigid-body docking by sampling billions of conformations;

(ii) root-mean-square deviation (RMSD)-based clustering of the 1,000 lowest-energy structures generated, to find the largest clusters that will represent the most likely models of the complex; and

9) **Validating Docked Structures:** The docked structures were again validated for their stereochemical properties. The Ramachandran plot is used to evaluate structures and to find whether the main chain torsion angles phi-psi (φ,ψ) torsion angles for all residues in the structure (except those at the chain termini) are stereochemically feasible. The different regions on the Ramachandran plot are as described in Morris et al. (1992) are as follows:

10) **Finding the interaction Profile:** The interaction profile was found using PDBSUM (http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/Generate.html) The PDBsum is a pictorial database that provides an overview of the contents of each 3D structure deposited in the Protein Data Bank (PDB).It shows molecule(s) that make up the structure (i.e. ligands, DNA, protein chains, and metal ions) and schematic diagrams of their interactions.

11) **Design the wash process accordingly:** Once the interaction profiles are known wash process could be focused on the type of interaction profile they are having and instead of using the storage buffer a separate wash buffer could be designed accordingly in order to break the bonds which are being formed between the HCPs and mAb.

RESULTS AND DISCUSSION

1) Retrieval of CHO Proteome

CHO proteome consists of total 23884 proteins was retrieved from its database. Out of which 3474 proteins had unknown amino acids in their sequences. The UniProt Proteome Id for (Cricetulus griseus) is UP000001075, it is a reference proteome it has been (manually and algorithmically) selected. It covers well-studied model organisms and other organisms of interest for pharmaceuticals, biomedical research and phylogeny.

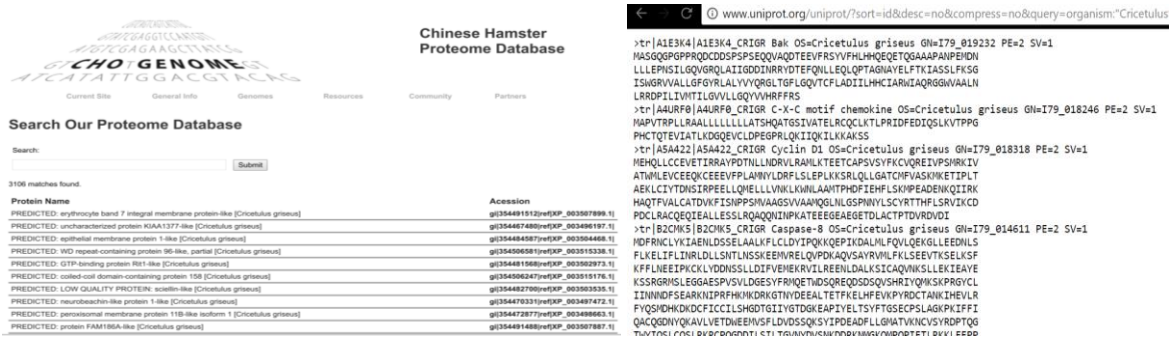


Figure 7: CHO proteome from the database in FASTA format

2) Literature search for the reported HCP's and their classification

Large number of wet lab studies have reported different HCP's. According to the type of studies they could be classified into three categories.

46 elements included exclusively in "Varying Expression"
78kDa glucose-regulated protein
Acid ceramidase
Beta 2-microglobulin
Cathepsin Z
Chondroitin sulfate proteoglycan 4
Cofilin-1
Collagen alpha-1(III) chain
Glutathione transferase class pi
Heat shock protein HSP 90-beta
Insulin-like growth factor-binding protein 4
kDa glucose-regulated protein
Lysosomal alpha-glucosidase
Lysyl oxidase-like 1
N(4)-(beta-N-acetylglucosaminyl)-L-asparaginase
Neural cell adhesion molecule 1
Nucleobindin-1
Nucleoside diphosphate kinase A
Nucleoside diphosphate kinase B
Peptidyl-prolyl cis-trans isomerase B
Peptidyl-prolyl cis-trans isomerase C
Peroxiredoxin-1
Pigment epithelium-derived factor
Putative phospholipase B-like 2
Serine protease
Thrombospondin-1
Vesicular integral-membrane protein VIP36
Latent TGF-beta complexed protein(LTCP)
Phospho lipid transfer protein
Extracellular matrix protein 1
Granulins
Fibronectin
Thrombospondin-3
Retinoid -inducible serine carboxypeptidase
Beta-galactosidases
Complement C3
Matrix metalloprotinase-9
Beta-glucuronidase
EMLIN-1
C-C motif chemokine 2
78 kDa glucose-regulated protein
Semaphorin-3E
Glucosidase 2 subunit beta
Hypoxia up-regulated protein 1
Neural alpha-glucosidase AB
Endoplasmin
Plasminogen activator inhibitor 1

24 elements included exclusively in "co elute"
Alpha-galactosidase A
Alpha-N-acetylgalactosaminidase
Amyloid-like protein 2
Calsyntenin-1
Dickkopf-related protein 3
Di-N-acetylchitobiase
dnaK-type molecular chaperone GRP78 precursor
Follistatin-related protein 1
Glucose regulated protein
Glyceraldehyde-3-phosphate dehydrogenase
Glypican-1
Histone H2A type 1
Leukemia inhibitory factor
MAM domain-containing protein 2
Matrix metalloproteinase-9
N-acetylglucosamine-6-sulfatase
N-sulphoglucosamine sulphohydrolase
Peptidyl-glycine alpha-amidating monooxygenase B
Procollagen-lysine,2-oxoglutarate 5-Dioxygenase 1
Ribonuclease T2
Semaphorin-3C
Syndecan-4
Tissue alpha-L-fucosidase
Tubulointerstitial nephritis antigen-like

8 elements included exclusively in "Product Associated"
ATP synthase subunit beta,mitochondrial
Complement C1r-a sub-component
Heat shock protein
Insulin
Neural cell adhesion molecule
Peroxirodoxin 1
Renin receptor
Vimentin

Figure 8: Experimentally verified HCPs categorized into three different categories of Product associated, coeluting, Varying Expression based on their study.

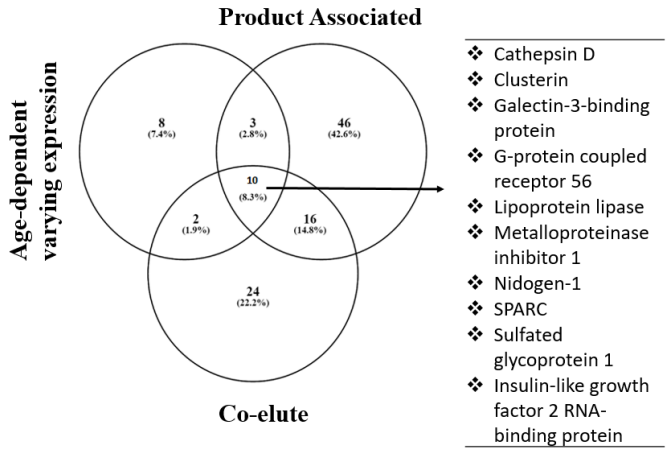


Figure 9: List of 10 HCPs selected from the experimentally verified based on this Venn diagram, the HCP's coexisting in all three categories are most difficult to remove HCPs and thus are further studied for their interactions

Finally, these 10 HCPs were taken for further study.

3) Hotspot Identification

The hotspots were identified using Aggrescan. Out of total 23884 proteins, 3474 proteins had unknown sequences and were not accepted by the software, also 406 protein sequences were too large to process. Thus, a total of 20004 Proteins were analysed out of which it was found that the no of hotspots ranges from 1-86 hotspots and in experimentally verified proteins HCPs had 3-49 hotspots. Referring to experimentally verified range 18966 proteins lied in that range making it difficult to be analysed but the large number of hotspots verified that CHO HCP's possess large number of hotspots and are very likely to elute with mAb during purification.

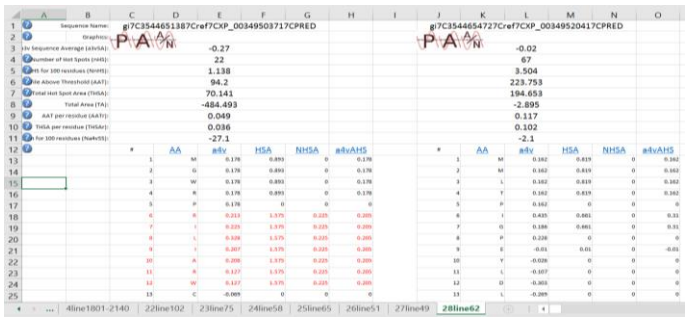


Figure 10: Results of Hotspots identified for the CHO Proteome using Aggrescan Software

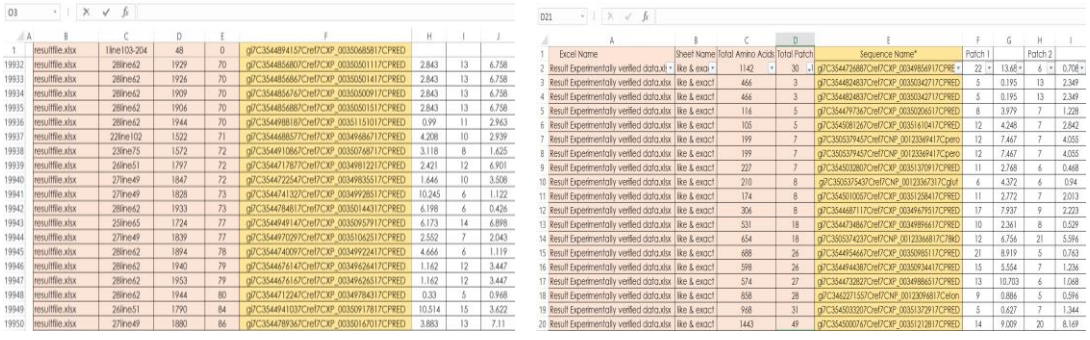


Figure 11: Aggrescan result summary for 18966 proteins

4) Structure Prediction

Models were constructed using different software's to get the best models out of all. Different software uses different modeling methods and algorithms to construct best models.All these softwares use different algorithms to generate these models.

i. **Bhageerath Results:** It gives 5 best models using ab initio modeling method.

a) Cathepsin D:

Figure 12: Cathepsin D structure wasn't modelled due to error in reading file

b) Clustrin:

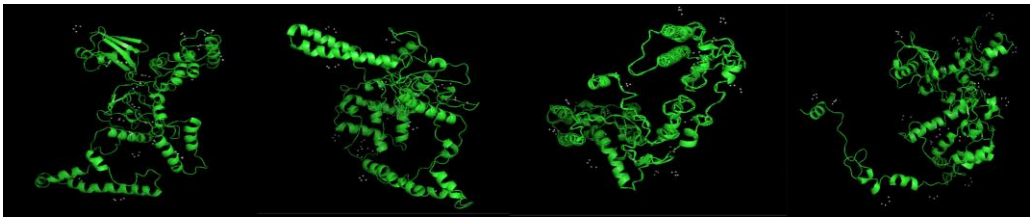


Figure 13: Only four structures are modelled for Clustrin

c) SPARC

Figure 14: SPARC structure wasn't modelled due to error in reading file

d) Lipoprotein lipase

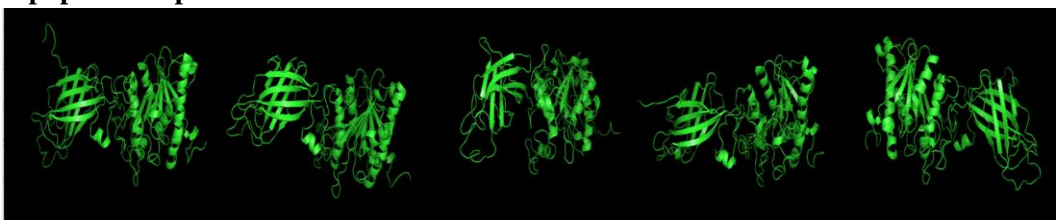


Figure 15: Five different structures were modelled for Lipoprotein lipase

e) Nidogen-1

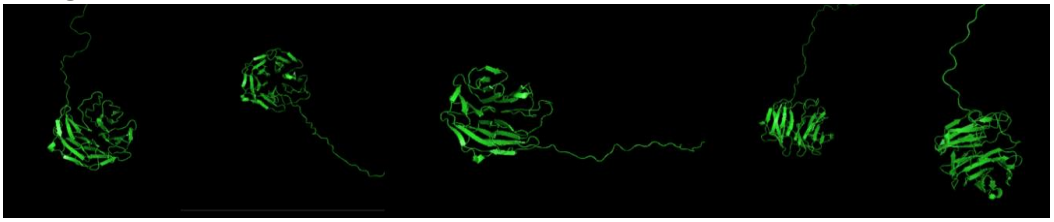


Figure 16: Five different structures were modelled for Nidogen-1

f) Metalloproteinase inhibitor 1

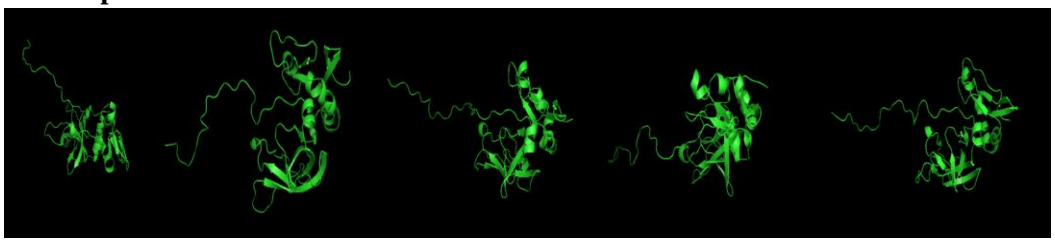


Figure 17: Five different structures were modelled for Metalloproteinase inhibitor 1

g) Sulfated glycoprotein 1

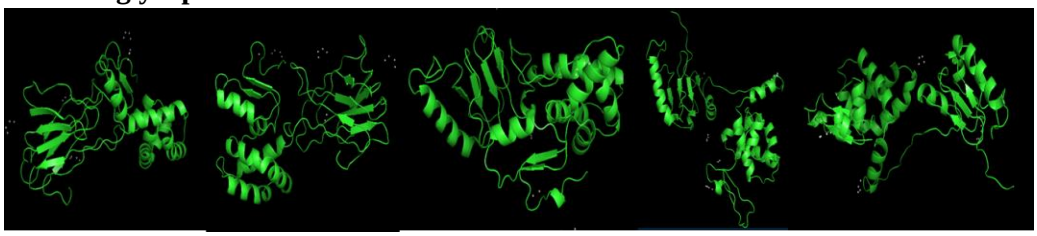


Figure 18: Five different structures were modelled for Sulfated glycoprotein 1

h) Insulin-like growth factor 2 RNA-binding protein

Figure 19: Insulin-like growth factor 2 RNA-binding protein structure wasn't modelled due to error in reading file

i) G protein coupled receptor

Figure 20: G protein coupled receptor structure wasn't modelled due to error in reading file

j) Galectin-3-binding protein

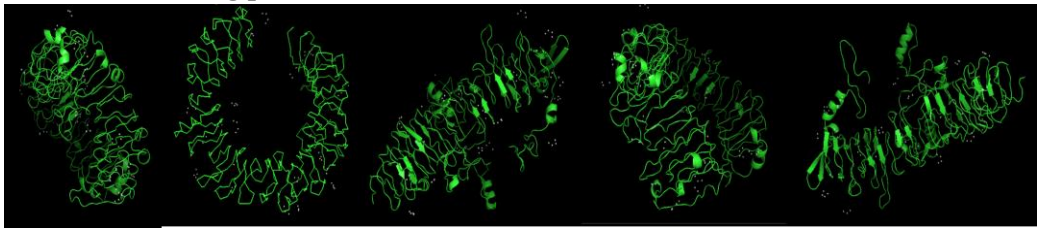


Figure 21: Five different structures were modelled for Galectin-3-binding protein

ii. **I-Tasser Results:** It generates 3-d atomic models from iterative structural assembly simulations and multiple threading alignments.

a) Cathepsin D

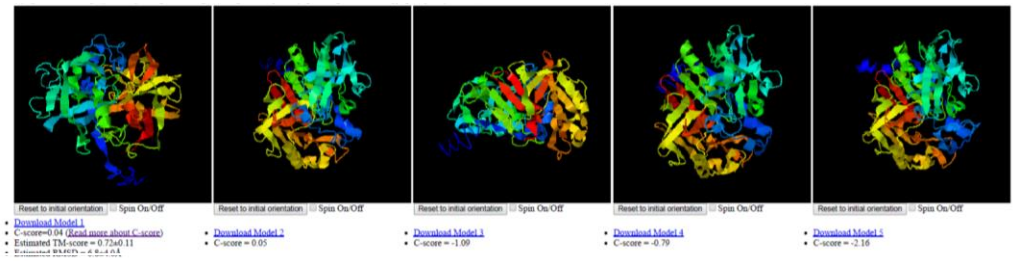


Figure 22: Five different structures were modelled for Cathepsin D

b) Clustrin

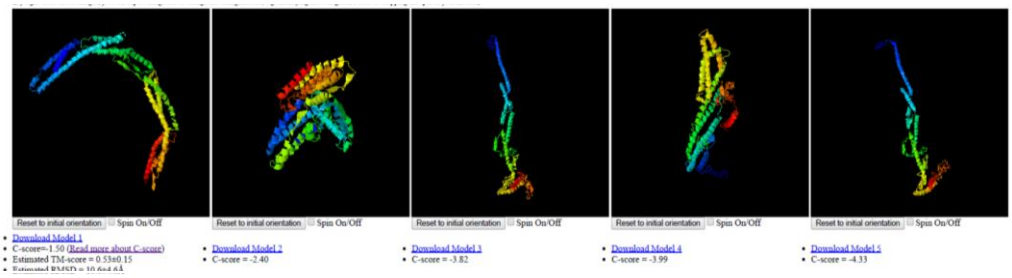


Figure 23: Five different structures were modelled for Clustrin

c) Galectin-3-binding protein

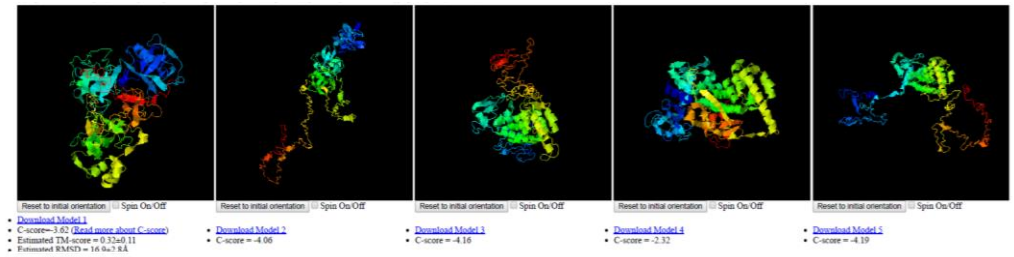


Figure 24: Five different structures were modelled for Galectin-3-binding protein

d) G protein coupled receptor

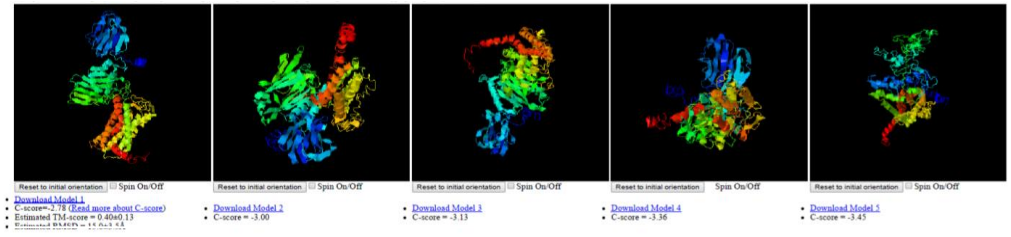


Figure 25: Five different structures were modelled for G-protein coupled receptor

e) Insulin-like growth factor 2 RNA-binding protein

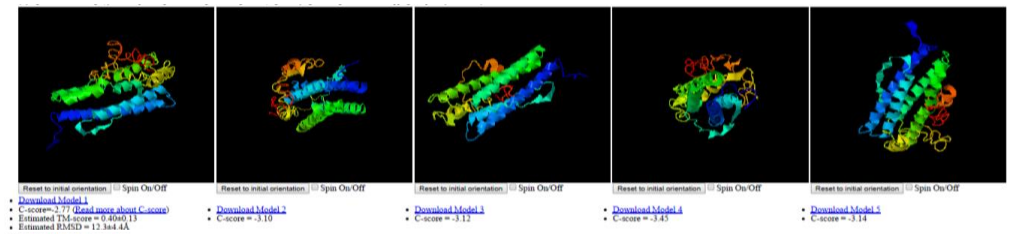


Figure 26: Five different structures were modelled for Insulin-like growth factor

f) Lipoprotein lipase

Figure 27: Only one structure was modelled for Lipoprotein lipase

g) Metalloproteinase inhibitor 1

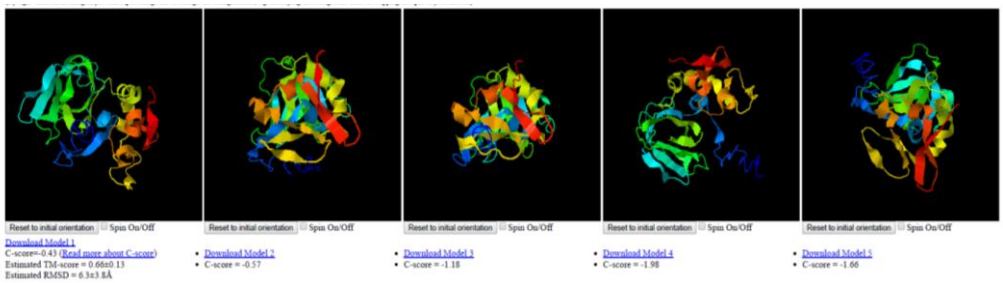


Figure 28: Five different structures were modelled for Metalloproteinase inhibitor 1

h) Nidogen-1

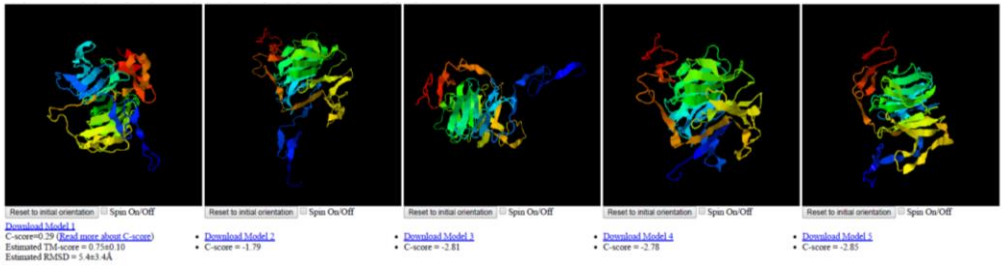


Figure 29: Five different structures were modelled for Nidogen-1

i) SPARC

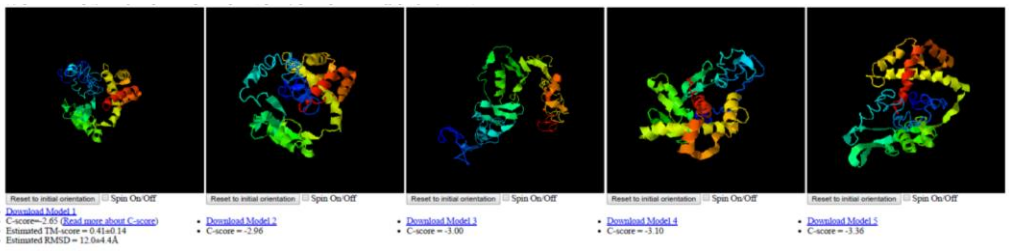


Figure 30: Five different structures were modelled for SPARC

j) Sulfated glycoprotein 1

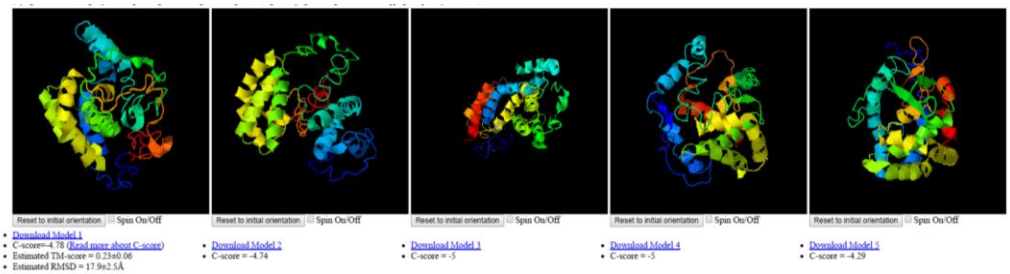


Figure 31: Five different structures were modelled for Insulin-like growth factor

iii. **Robetta Results:** It gives 5 best models using de novo or comparative modeling structure prediction methods. Also, it can be used for modeling individual chain structures if needed.

a. Cathepsin D

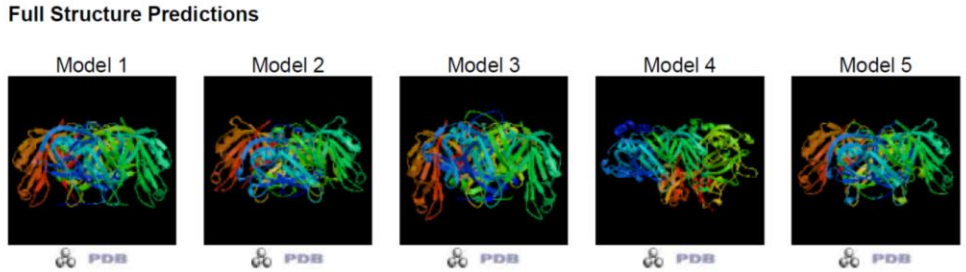


Figure 32: Five different structures were modelled for Cathepsin D

b. Clustrin: Results not predicted

c. Galectin-3-binding protein: Results not predicted

d. G protein coupled receptor

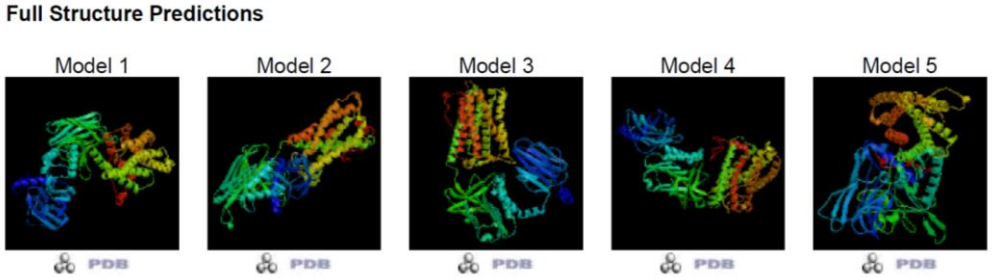


Figure 33: Five different structures were modelled for G protein coupled receptor

e. Insulin-like growth factor 2 RNA-binding protein

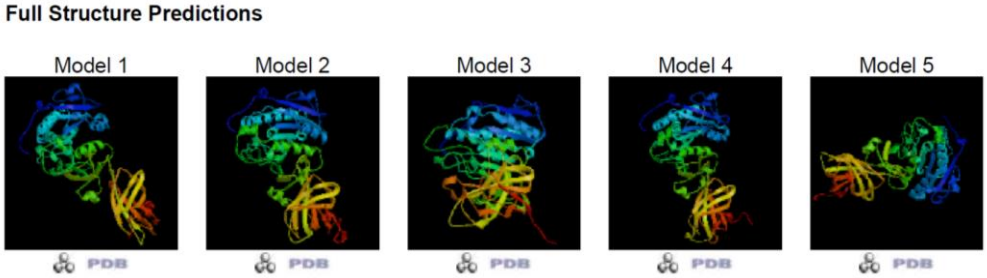


Figure 34: Five different structures were modelled for Insulin-like growth factor

f. Lipoprotein lipase

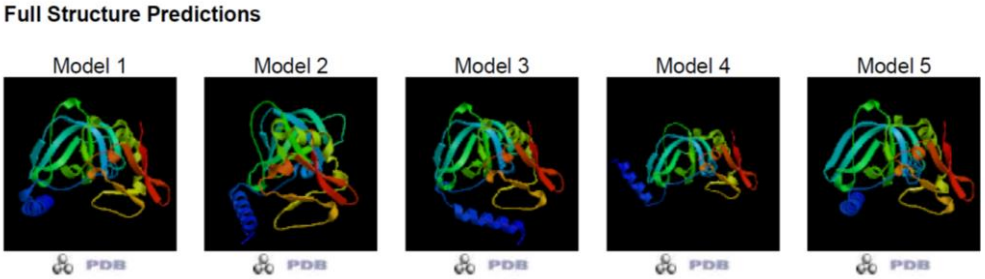


Figure 35: Five different structures were modelled for Lipoprotein lipase

g. Metalloproteinase inhibitor 1

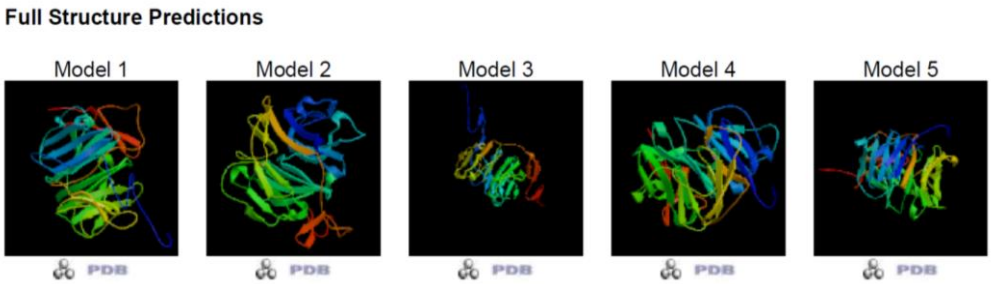


Figure 36: Five different structures were modelled for Metalloproteinase inhibitor 1

h. Nidogen-1

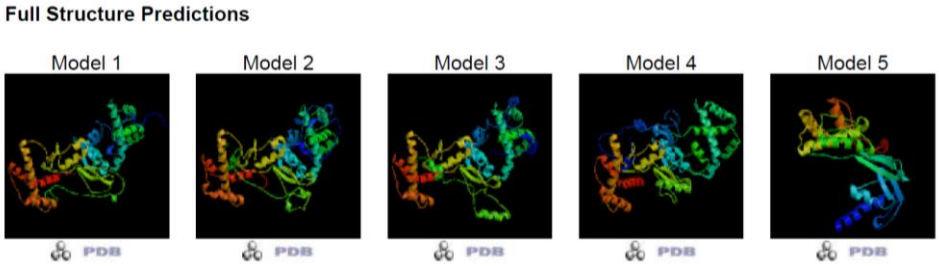


Figure 37: Five different structures were modelled for Nidogen-1

i. SPARC

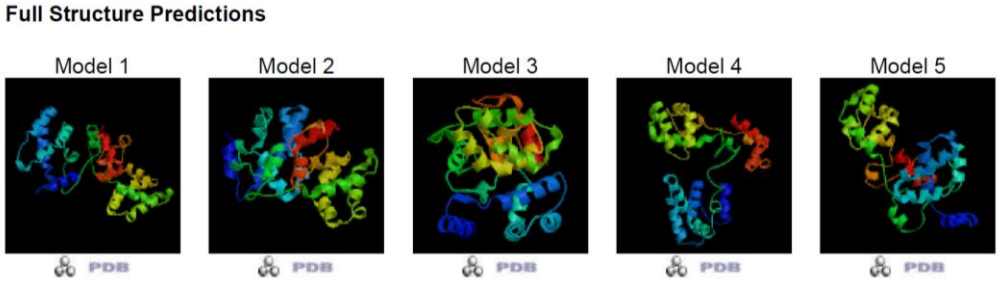


Figure 38: Five different structures were modelled for SPARC

j. Sulfated glycoprotein 1

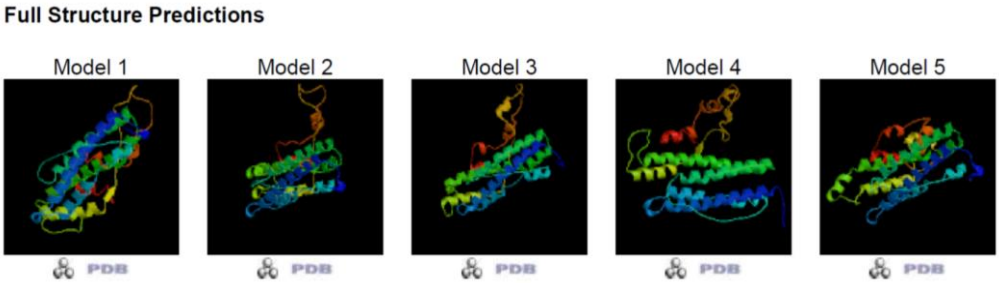


Figure 39: Five different structures were modelled for Sulfated glycoprotein 1

iv. Multicom Raptor-X Results: It uses template-based tertiary structure modeling, and probabilistic alignment sampling. It can also detect remotely related template sequence for a given sequence by a novel nonlinear context-specific probabilistic consistency algorithm and alignment potential

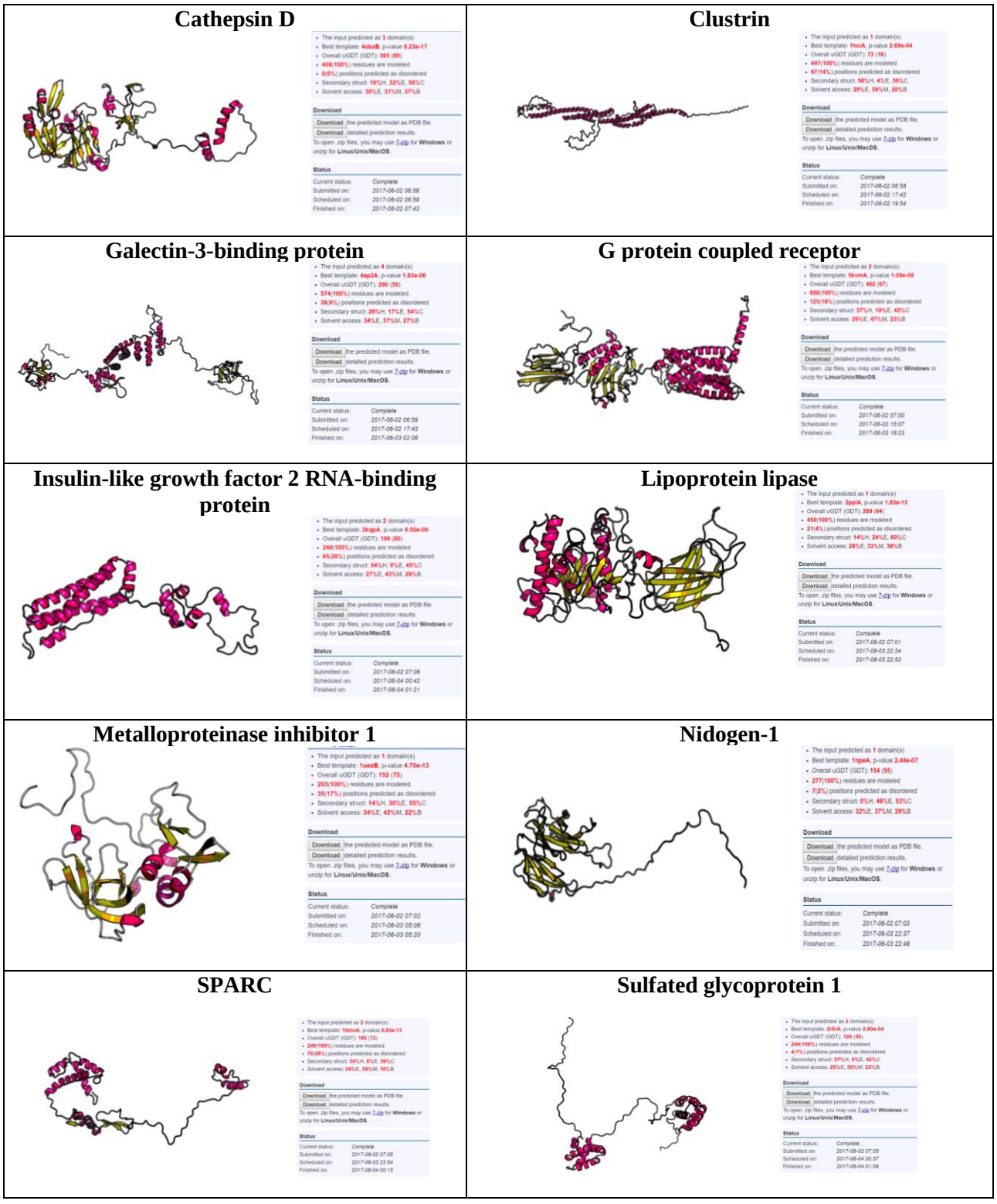


Figure 40: Raptor-X generates one best model for each protein using homology detection methods

v. **Phyre 2 Results:** Phyre2, uses advanced remote homology detection methods to build 3D models, and to analyse the effect of amino acid variants and predict ligand binding sites for a given protein sequence.

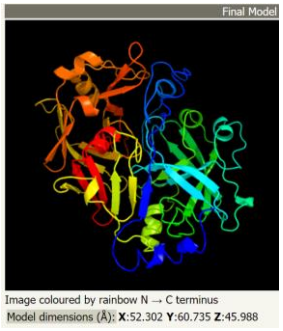
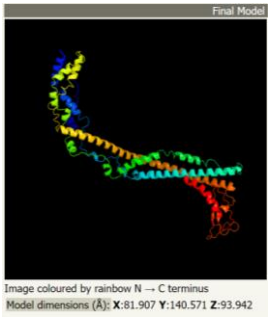
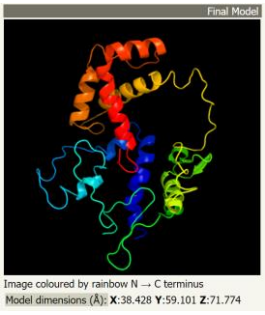
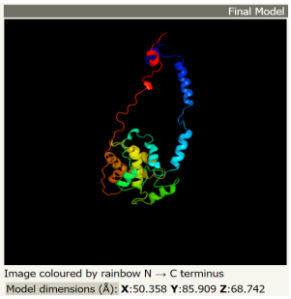
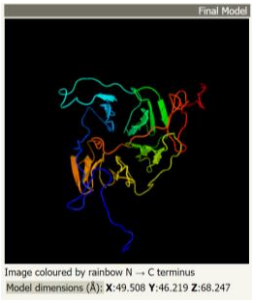
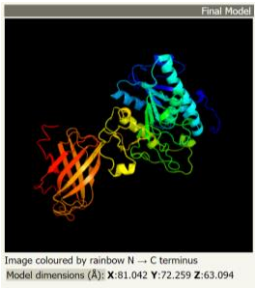
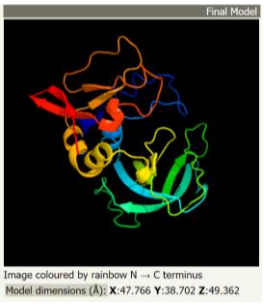
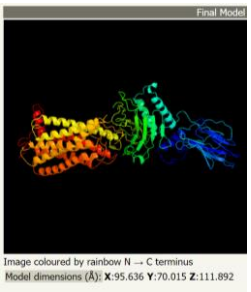
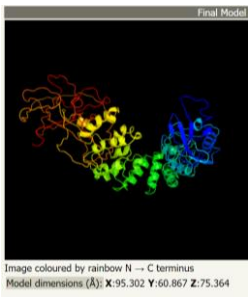
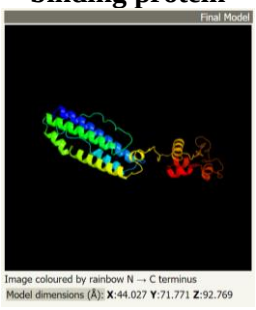
Cathepsin D	**Clustrin**	**SPARC**	**Sulfated glycoprotein 1**
			
Nidogen-1	**Lipoprotein lipase**	**Metalloproteinase inhibitor 1**	**G protein coupled receptor**
			
Galectin-3-binding protein	**Insulin-like growth factor 2 RNA-binding protein**		
			

Figure 41: Phyre 2 generates one best model for each protein using homology detection methods

5) Structure Validation

Total of 151 models were found which were further validated using ProTSAV and SAVES metaservers. ProTSAV gives the score for the models using different tools analysis online. While SAVES server gives the graphs, scores and errors so as to identify the best models.

Representative ProTSAV Result: ProTSAV gives the scores for all structures by after validating them using different individual servers. Scores below 0.4 are considered good.

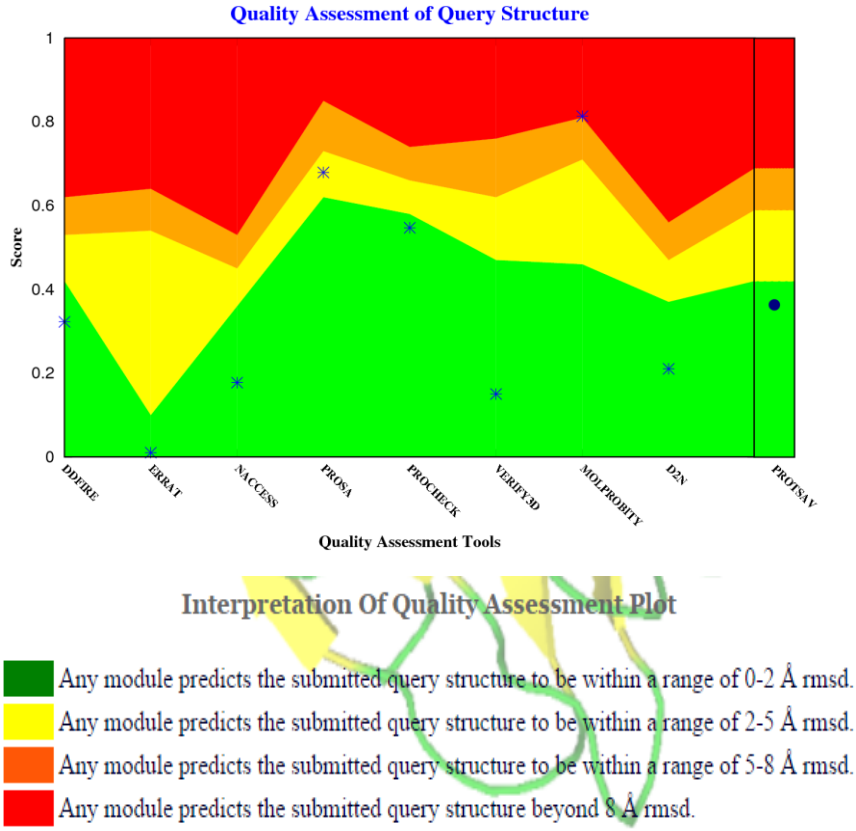


Figure 42: ProTSAV gives the graph with score, the score lying in green region are considered the best.

Representative SAVES Result

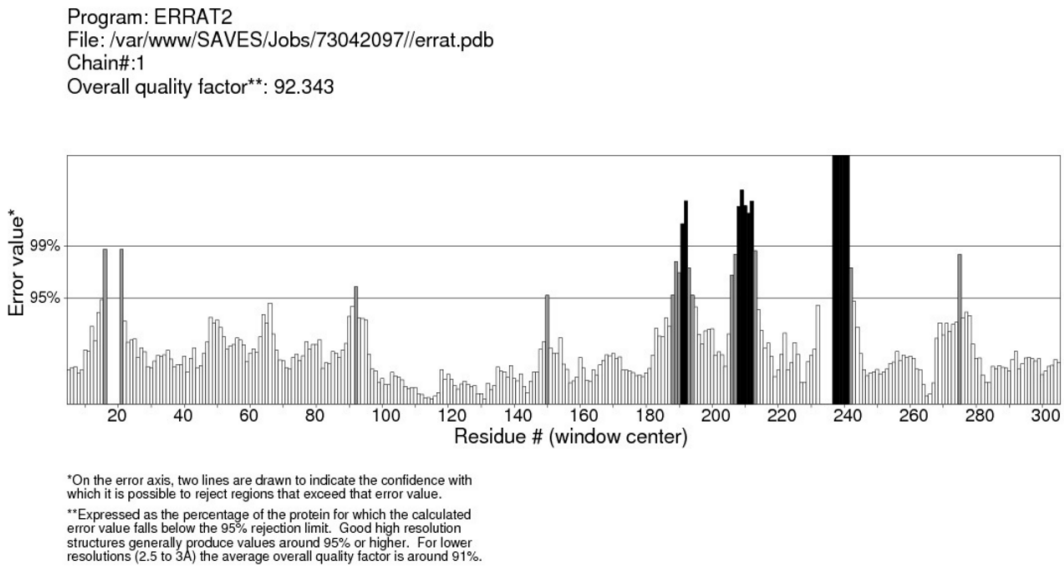


Figure 43: Errat on SAVES server gives the overall quality factor, best scores is around 95% or higher for lower resolution quality factor could be around 91%.

Representative Verify 3D Result

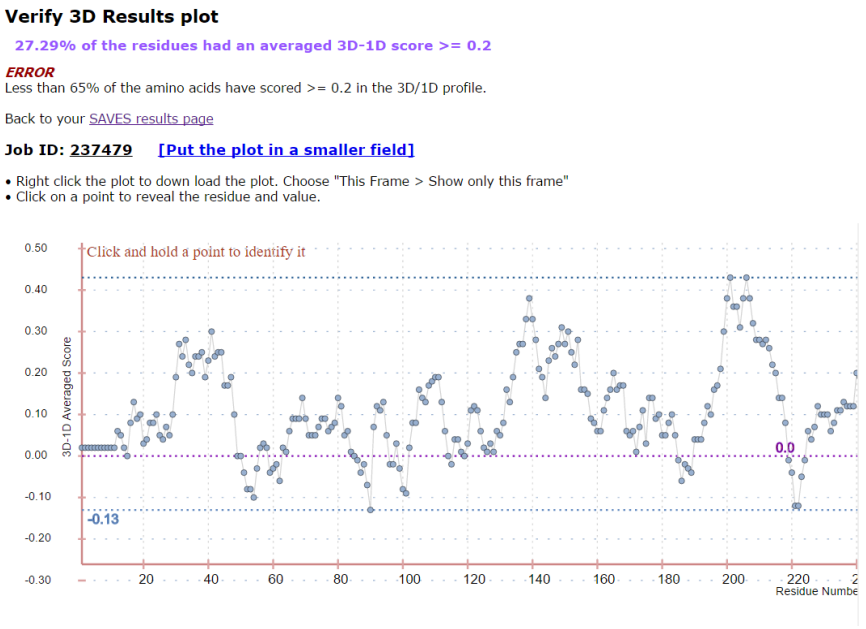


Figure 44: Verfy 3D gives the graph with scores, in which above 65% of amino acids scoring above 0.2 in 3D/1D profile is considered acceptable.

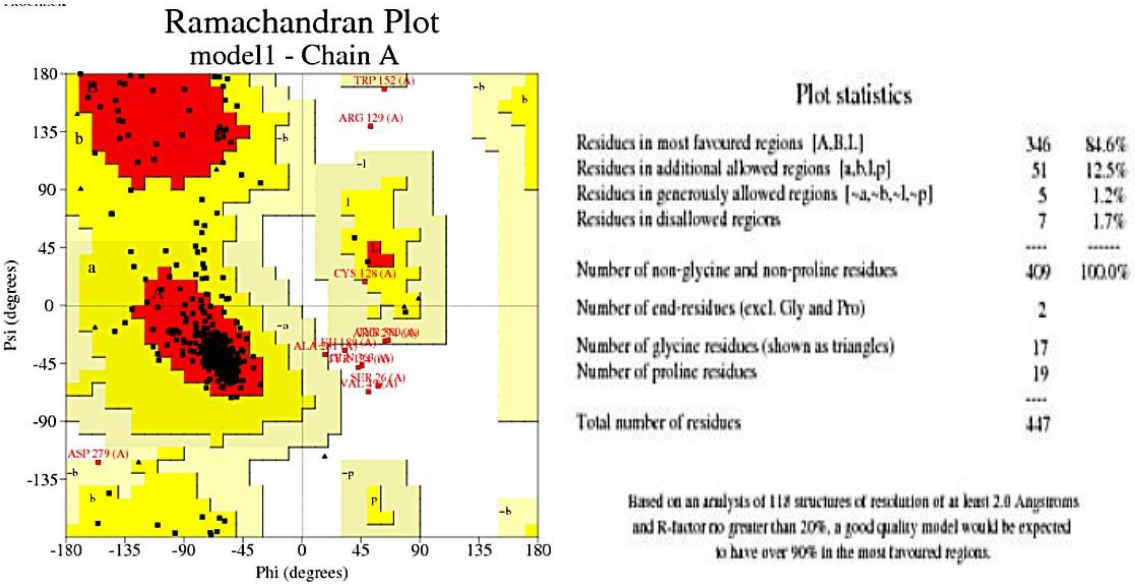


Figure 45: Phyre 2 generates one best model for each protein using homology detection methods

Most favoured regions	[A,B,L]	452	87.1%**
Additional allowed regions	[a,b,l,p]	118	10.1%
Generously allowed regions	[~a,~b,~l,~p]	8	1.4%
Disallowed regions	[XX]	8	1.4%*
Non-glycine and non-proline residues		586	100.0%
End-residues (excl. Gly and Pro)		5	
Glycine residues		52	
Proline residues		36	
Total number of residues		679	

Ideally most favoured regions in a Ramachandran plot should be 90% but generally >80% is considered good if the total of most favoured regions and additionally allowed regions comes to be 90% or above.

6) Energy Minimization

Out of 151 models which were validated using ProTSAV and SAVES metaservers and it was found that their scores were low so the models with highest scores were taken to be refined using 3D Refine and Galaxy Refine.

i. 3D Refine:

3-D Refine	I-Taseer	Bhageerath H	Phyre-2	Multicom Raptor-X	Robetta
Cathepsin D	✓	✓	✓	✓	Error multiple chains reported
CLUSTRIN	✓	✓	✓	✓	Error multiple chains reported
Galectin-3-binding protein	✓	✓	✓	✓	Error multiple chains reported
Gprotein coupled	✓	Error! Invalid Input Format	✓	✓	Error multiple chains reported
Insulin	✓	Error! Invalid Input Format	✓	✓	Error multiple chains reported
Lipoprotein lipase	✓	✓	✓	✓	Error multiple chains reported
Metalloproteinase inhibitor 1	✓	✓	✓	✓	Error multiple chains reported
Nidogen-1	✓	✓	✓	✓	Error multiple chains reported
SPARC	✓	Error! Invalid Input Format	✓	✓	Error multiple chains reported
Sulfated glycoprotein 1	✓	✓	✓	✓	Error multiple chains reported

Figure 46: 3-D refine accepted all the models except few Bhageerath H and Robetta models, later few models gave error of missing residues in further analysis.

ii. Galaxy Refine

Galaxy Refine	I-Taseer	Bhageerath H	Phyre-2	Multicom Raptor-X	Robetta
Cathepsin D	✓	✓	✓	✓	The input PDB file contains multiple chains.
CLUSTRIN	✓	Non-standard residue in ATOM records.	Break in protein chain.	✓	The input PDB file contains multiple chains.
Galectin-3-binding protein	✓	Non-standard residue in ATOM records.	Break in protein chain.	Break in protein chain.	The input PDB file contains multiple chains.
Gprotein coupled	✓	Error! Invalid Input Format	✓	✓	The input PDB file contains multiple chains.
Insulin	✓	Error! Invalid Input Format	Break in protein chain.	✓	The input PDB file contains multiple chains.
Lipoprotein lipase	✓	✓	✓	✓	The input PDB file contains multiple chains.
Metalloproteinase inhibitor 1	✓	✓	✓	✓	The input PDB file contains multiple chains.
Nidogen-1	✓	✓	✓	✓	The input PDB file contains multiple chains.
SPARC	✓	Error! Invalid Input Format	✓	✓	The input PDB file contains multiple chains.
Sulfated glycoprotein 1	✓	Non-standard residue in ATOM records.	✓	✓	The input PDB file contains multiple chains.

Figure 47: Galaxy refine showed error models in Bhageerath H, Phyre2, Raptor-X and Robetta models analysis.

Comparing both results it is evident that I-Tasser models are most well-structured models as all models are accepted for refining while models generated using other tools showed various errors during validation also (missing residues error) Thus, I-Tasser models are chosen for further work.

7) **Structure Validation:** Best structures were validated and were used for further analysis. It was found that galaxy refine gave the least ProTSAV score for most of the models except Nidogen-1 whose initial structure had the least score thus taken as it is.

I-Tasser Model	Cathepsin D	CLUSTRIN	Galectin-3-binding protein	G protein coupled receptor	Insulin	Lipoprotein lipase	Metalloproteinase inhibitor 1	Nidogen-1	SPARC	Sulfated glycoprotein 1
Model	Model 4	Model 2	Model 4	Model 2	Model 1	Model 1	Model 3	Model 1	Model 5	Model 1
ProTsav Score before refining	0.553	0.6391	0.6092	0.7054	0.628	0.6261	0.5878	0.5616	0.695	0.9617
3-D Refined	Model 3	Model 5	Model 5	Model 3	Model 5	Model 5	Model 5	Model 5	Model 5	Model 5
Score	0.6354	0.7286	0.6991	0.6969	0.6855	0.6334	0.6688	0.7203	0.7896	0.7398
Galaxy Refine	Model 3	Model 3	Model 3	Model 2	Model 5	Model 5	Model 2	Model 5	Model 2	Model 4
Score	0.5236	0.6025	0.5763	0.6134	0.5994	0.5103	0.573	0.6102	0.5843	0.6702

Figure 48: Galaxy refine gave the lowest scores during revalidation, except Nidogen-1 whose initial score was best, thus models with lowest scores were taken for Docking with mAb.

Best Structures were selected for final docking with the mAb.

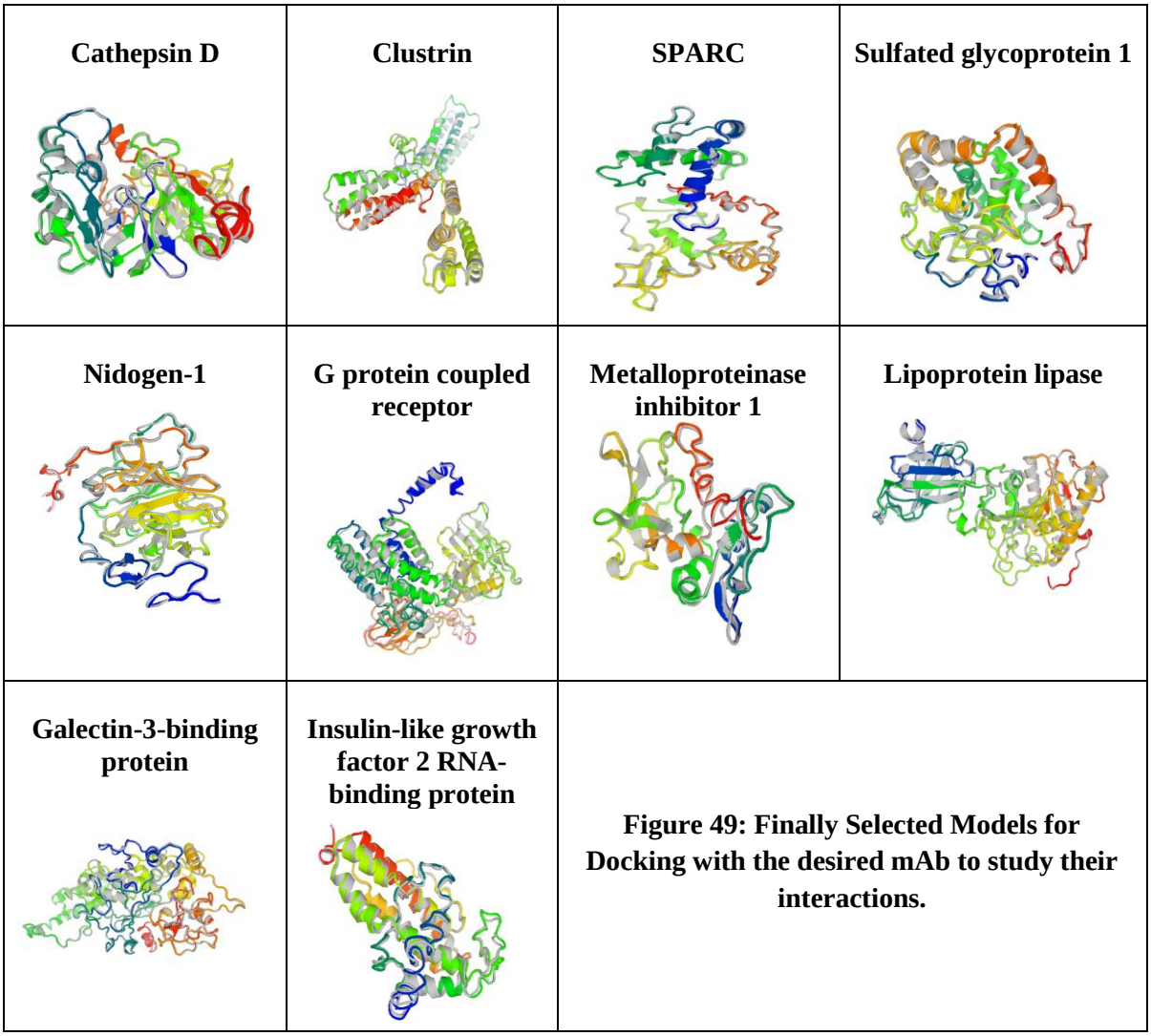
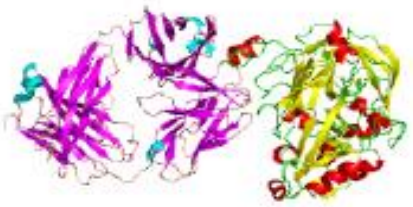


Figure 49: Finally Selected Models for Docking with the desired mAb to study their interactions.

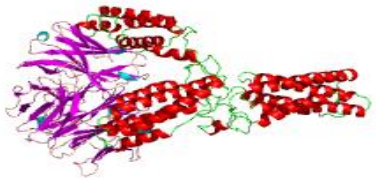
8) **Docking:** Docking was done for the best predicted models with the Rituximab (PDB ID 4KAQ) and the stable structures were obtained using Cluspro. It docks structures in three categories balanced, hydrophobic, electrostatic and Vdw+elec and gives around 30 best docked structures for each at different angles. Out of which the least lowest energy scores were chosen from balanced structure category to study the interaction profiles.

Cathepsin D



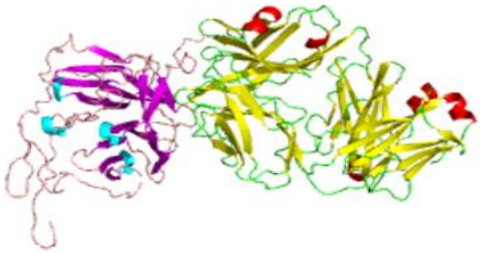
Cluster	Members	Representative	Weighted Score
0	87	Center	-644.2
		Lowest Energy	-678.7
1	70	Center	-695.6
		Lowest Energy	-751.4
2	62	Center	-671.2
		Lowest Energy	-691.2

Clustrin



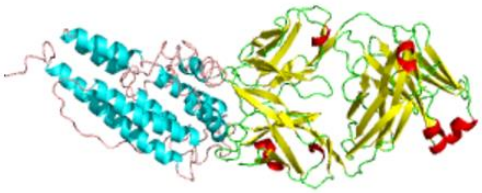
Cluster	Members	Representative	Weighted Score
0	74	Center	-789.0
		Lowest Energy	-988.7
1	69	Center	-933.5
		Lowest Energy	-1081.9
2	57	Center	-870.9
		Lowest Energy	-952.3

Nidogen-1



Cluster	Members	Representative	Weighted Score
0	67	Center	-718.9
		Lowest Energy	-718.9
1	66	Center	-862.8
		Lowest Energy	-866.2
2	58	Center	-661.6
		Lowest Energy	-808.7

Insulin-like growth factor 2 RNA-binding protein

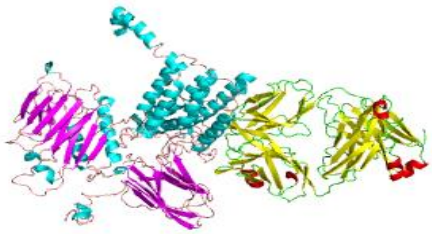


Cluster	Members	Representative	Weighted Score
0	83	Center	-958.0
		Lowest Energy	-958.0
1	82	Center	-727.4
		Lowest Energy	-867.5
2	62	Center	-840.2
		Lowest Energy	-945.9

Galectin-3-binding protein

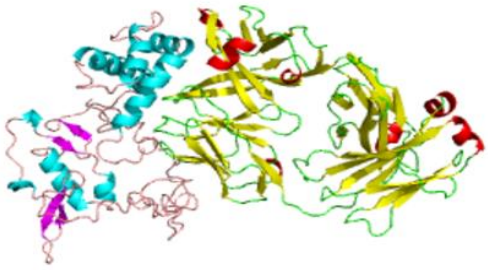
Cluster	Members	Representative	Weighted Score
0	141	Center	-759.8
		Lowest Energy	-957.8
1	63	Center	-866.0
		Lowest Energy	-866.0
2	59	Center	-746.9
		Lowest Energy	-825.1

G protein coupled receptor



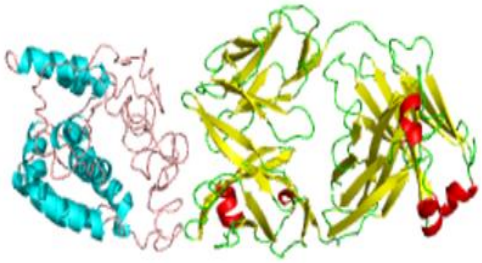
Cluster	Members	Representative	Weighted Score
0	100	Center	-933.9
		Lowest Energy	-1115.1
1	44	Center	-948.9
		Lowest Energy	-1030.8
2	43	Center	-952.3
		Lowest Energy	-972.4

SPARC



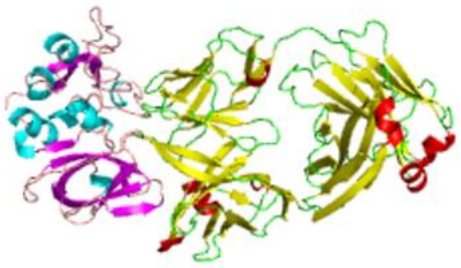
Cluster	Members	Representative	Weighted Score
0	83	Center	-796.2
		Lowest Energy	-892.7
1	80	Center	-660.1
		Lowest Energy	-846.7
2	77	Center	-701.4
		Lowest Energy	-887.7

Sulfated glycoprotein 1



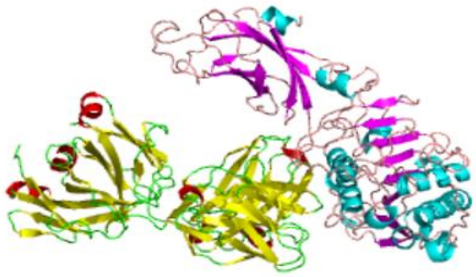
Cluster	Members	Representative	Weighted Score
0	77	Center	-550.4
		Lowest Energy	-656.8
1	71	Center	-762.7
		Lowest Energy	-762.7
2	55	Center	-602.4
		Lowest Energy	-620.4

Metalloproteinase inhibitor 1



Cluster	Members	Representative	Weighted Score
0	170	Center	-746.3
		Lowest Energy	-937.1
1	107	Center	-704.7
		Lowest Energy	-793.4
2	80	Center	-768.6
		Lowest Energy	-897.0

Lipoprotein lipase



Cluster	Members	Representative	Weighted Score
0	84	Center	-639.9
		Lowest Energy	-705.6
1	45	Center	-618.0
		Lowest Energy	-707.0
2	40	Center	-638.9
		Lowest Energy	-797.1

Figure 50: Docked models using Cluspro with lowest scores in balanced category were selected to study the interaction between them

Validating Docked Structures: The docked structures were again validated for their stereochemical properties. The Ramachandran plot is used to evaluate structures and to find whether the main chain torsion angles phi-psi (φ,ψ) torsion angles for all residues in the structure (except those at the chain termini) are stereochemically feasible. The different regions on the Ramachandran plot are as described in Morris et al. (1992) are as follows:

Most favoured regions are indicated by-

A - Core alpha

B - Core beta

L - Core left-handed alpha

Additional allowed regions are indicated by

a - Allowed alpha

b - Allowed beta

l - Allowed left-handed alpha

p - Allowed epsilon

Generously allowed regions are indicated by

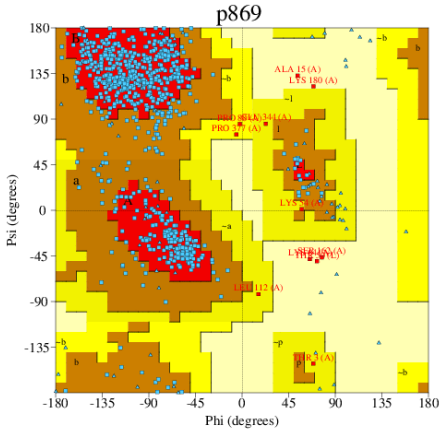
~p - Generous epsilon

~b - Generous beta

~a - Generous alpha

~l - Generous left-handed alpha

1. Cathepsin D & Rituximab

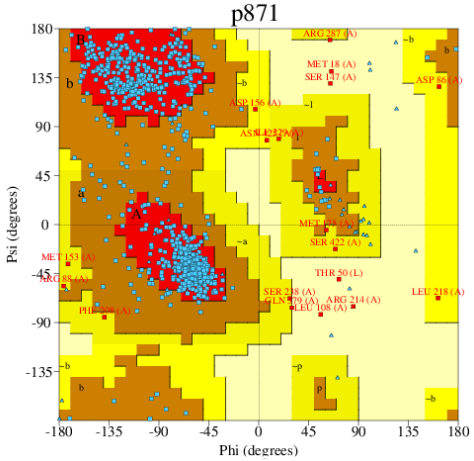


1. Ramachandran Plot statistics

		No. of residues	%-tage
Most favoured regions	[A,B,L]	577	81.7%*
Additional allowed regions	[a,b,l,p]	120	17.0%
Generously allowed regions	[~a,~b,~l,~p]	4	0.6%
Disallowed regions	[XX]	5	0.7%*
Non-glycine and non-proline residues		706	100.0%
End-residues (excl. Gly and Pro)		5	
Glycine residues		79	
Proline residues		49	
Total number of residues		839	

Figure 51: Docked structures were validated and it was found that 81.7% residues in most favoured region and 17.0% in additionally allowed region making it 98%

2. Clustrin & Rituximab

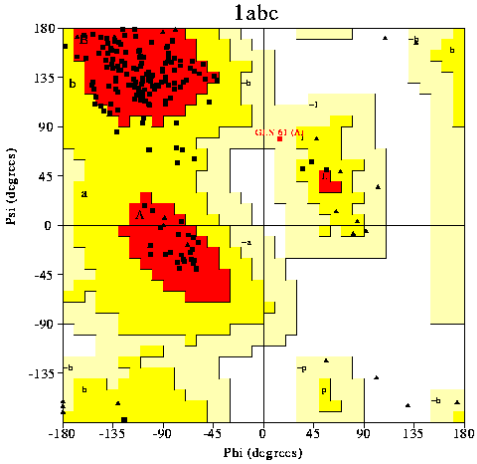


1. Ramachandran Plot statistics

		No. of residues	%-tage
Most favoured regions	[A,B,L]	662	85.2%*
Additional allowed regions	[a,b,l,p]	97	12.5%
Generously allowed regions	[~a,~b,~l,~p]	12	1.5%
Disallowed regions	[XX]	6	0.8%*
Non-glycine and non-proline residues		777	100.0%
End-residues (excl. Gly and Pro)		5	
Glycine residues		51	
Proline residues		45	
Total number of residues		878	

Figure 52: Docked structures were validated and it was found that 85.2% residues in most favoured region and 12.5% in additionally allowed region making it 97.7%

3. SPARC & Rituximab

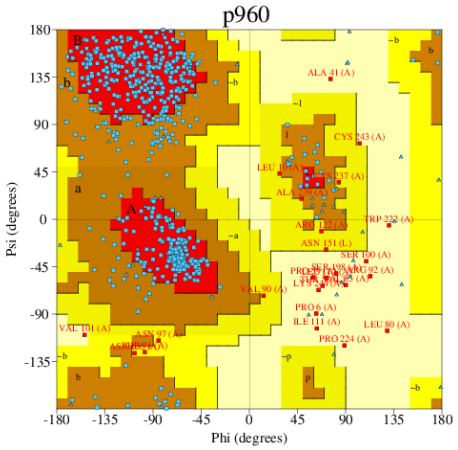


1. Ramachandran Plot statistics

		No. of residues	%-tage
Most favoured regions	[A,B,L]	457	78.0%*
Additional allowed regions	[a,b,l,p]	115	19.6%
Generously allowed regions	[~a,~b,~l,~p]	4	0.7%
Disallowed regions	[XX]	10	1.7%*
Non-glycine and non-proline residues		586	100.0%
End-residues (excl. Gly and Pro)		4	
Glycine residues		48	
Proline residues		42	
Total number of residues		680	

Figure 53: Docked structures were validated and it was found that 78% residues in most favoured region and 19.6% in additionally allowed region making it 97.6%

4. Sulfated glycoprotein 1 & Rituximab

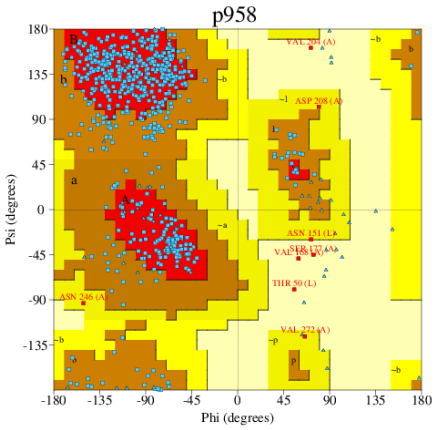


1. Ramachandran Plot statistics

		No. of residues	%-tage
Most favoured regions	[A,B,L]	450	76.5%**
Additional allowed regions	[a,b,l,p]	116	19.7%
Generously allowed regions	[~a,~b,~l,~p]	9	1.5%
Disallowed regions	[XX]	13	2.2%*
Non-glycine and non-proline residues		588	100.0%
End-residues (excl. Gly and Pro)		5	
Glycine residues		44	
Proline residues		43	
Total number of residues		680	

Figure 54: Docked structures were validated and it was found that 76.5% residues in most favoured region and 19.7% in additionally allowed region making it 96.2%

5. Nidogen-1 & Rituximab

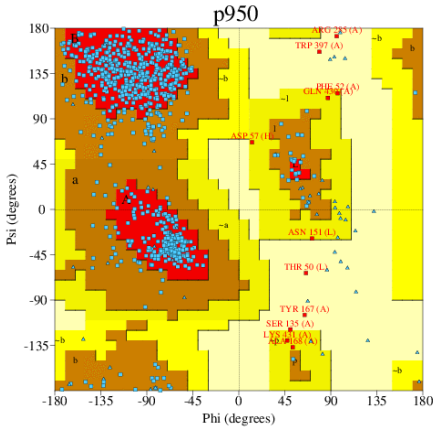


1. Ramachandran Plot statistics

		No. of residues	%-tage
Most favoured regions	[A,B,L]	462	77.4%**
Additional allowed regions	[a,b,l,p]	127	21.3%
Generously allowed regions	[~a,~b,~l,~p]	4	0.7%
Disallowed regions	[XX]	4	0.7%*
Non-glycine and non-proline residues		597	100.0%
End-residues (excl. Gly and Pro)		5	
Glycine residues		61	
Proline residues		46	
Total number of residues		709	

Figure 55: Docked structures were validated and it was found that 77.4% residues in most favoured region and 21.3% in additionally allowed region making it 98.7%

6. Lipoprotein lipase & Rituximab

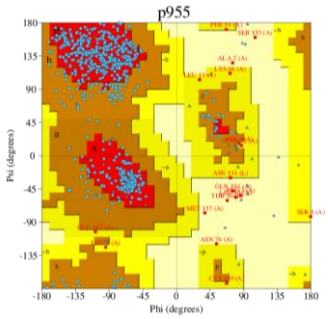


1. Ramachandran Plot statistics

		No. of residues	%-tage
Most favoured regions	[A,B,L]	630	82.2%*
Additional allowed regions	[a,b,l,p]	125	16.3%
Generously allowed regions	[~a,~b,~l,~p]	6	0.8%
Disallowed regions	[XX]	5	0.7%*
Non-glycine and non-proline residues		766	100.0%
End-residues (excl. Gly and Pro)		4	
Glycine residues		66	
Proline residues		45	
Total number of residues		881	

Figure 56: Docked structures were validated and it was found that 82.2% residues in most favoured region and 16.3% in additionally allowed region making it 98.5%

7. Metalloproteinase inhibitor 1 & Rituximab

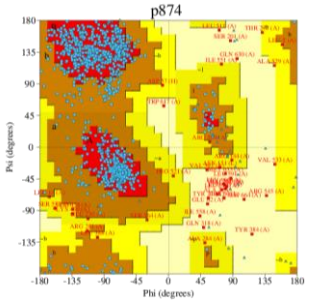


1. Ramachandran Plot statistics

		No. of residues	%-tage
Most favoured regions	[A,B,L]	432	79.3%**
Additional allowed regions	[a,b,l,p]	96	17.6%
Generously allowed regions	[~a,~b,~l,~p]	10	1.8%
Disallowed regions	[XX]	7	1.3%*
Non-glycine and non-proline residues		545	100.0%
End-residues (excl. Gly and Pro)		5	
Glycine residues		45	
Proline residues		39	
Total number of residues		634	

Figure 57: Docked structures were validated and it was found that 79.3% residues in most favoured region and 17.6% in additionally allowed region making it 96.9%

8. G protein coupled receptor & Rituximab

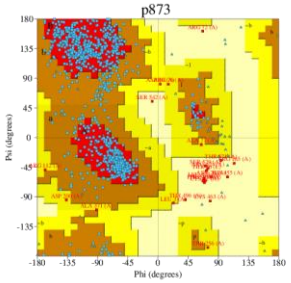


1. Ramachandran Plot statistics

		No. of residues	%-tage
Most favoured regions	[A,B,L]	789	79.7%**
Additional allowed regions	[a,b,l,p]	162	16.4%
Generously allowed regions	[~a,~b,~l,~p]	16	1.6%
Disallowed regions	[XX]	23	2.3%*
Non-glycine and non-proline residues		990	100.0%
End-residues (excl. Gly and Pro)		5	
Glycine residues		67	
Proline residues		57	
Total number of residues		1119	

Figure 58: Docked structures were validated and it was found that 79.7% residues in most favoured region and 16.4% in additionally allowed region making it 96.1%

9. Galectin-3-binding protein & Rituximab

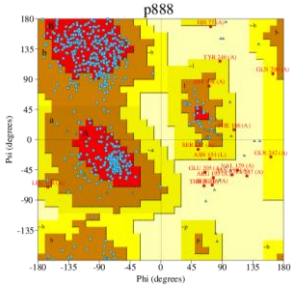


1. Ramachandran Plot statistics

		No. of residues	%-tage
Most favoured regions	[A,B,L]	689	78.7%**
Additional allowed regions	[a,b,l,p]	164	18.7%
Generously allowed regions	[~a,~b,~l,~p]	8	0.9%
Disallowed regions	[XX]	15	1.7%*
Non-glycine and non-proline residues		876	100.0%
End-residues (excl. Gly and Pro)		5	
Glycine residues		72	
Proline residues		52	
Total number of residues		1005	

Figure 59: Docked structures were validated and it was found that 78.7% residues in most favoured region and 18.7% in additionally allowed region making it 97.4%

10. Insulin-like growth factor 2 RNA-binding protein & Rituximab



1. Ramachandran Plot statistics

		No. of residues	%-tage
Most favoured regions	[A,B,L]	452	77.1%**
Additional allowed regions	[a,b,l,p]	118	20.1%
Generously allowed regions	[~a,~b,~l,~p]	8	1.4%
Disallowed regions	[XX]	8	1.4%*
Non-glycine and non-proline residues		586	100.0%
End-residues (excl. Gly and Pro)		5	
Glycine residues		52	
Proline residues		36	
Total number of residues		679	

Figure 60: Docked structures were validated and it was found that 77.1% residues in most favoured region and 20.1% in additionally allowed region making it 97.2%

9) **Finding the interaction Profile:** The interaction profile was found using PDB-SUM to identify the kind of bonds being formed between the protein chains of the docked structure, and it was found that mainly there were hydrogen bonds and salt bridges making the HCP's intact with the mAb thus, making them difficult to remove.

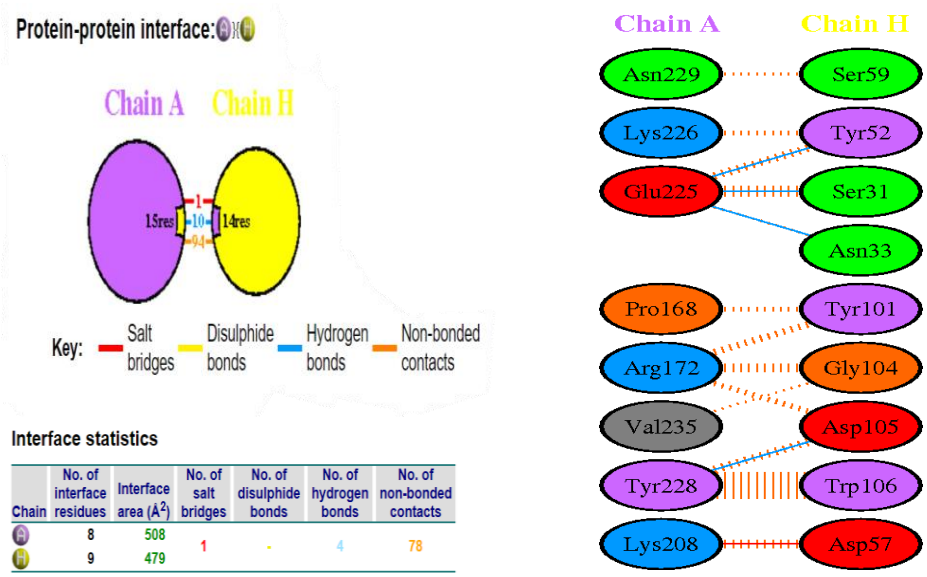


Figure 61: Representative Interaction Profile showing the number and kind of interactions between different protein-protein chains, out of Salt Bridges, Disulphide Bonds, Hydrogen Bonds, Non-Bonded contacts i.e., Vander wall forces etc. depicted by different color line segments.

Cathepsin

Chains	No. of interface residues	Interface area (Å²)	No. of salt bridges	No. of disulphide bonds	No. of hydrogen bonds	No. of non-bonded contacts
AH	9:12	571:530	-	-	2	56
AL	10:8	468:530	-	-	5	62
HL	35:38	1801:1829	1	-	9	197

Clustrin

Chains	No. of interface residues	Interface area (Å²)	No. of salt bridges	No. of disulphide bonds	No. of hydrogen bonds	No. of non-bonded contacts
AH	17:23	996:945	1	-	9	99
AL	9:11	616:587	1	-	4	62
HL	34:39	1835:1843	2	-	10	212

Insulin-like growth factor 2 RNA-binding protein

Chains	No. of interface residues	Interface area (Å²)	No. of salt bridges	No. of disulphide bonds	No. of hydrogen bonds	No. of non-bonded contacts
AH	23:22	1010:995	1	-	7	191
AL	7:5	236:274	-	-	3	23
HL	39:41	1964:1894	3	-	26	314

Galectin-3-binding protein

Chains	No. of interface residues	Interface area (Å²)	No. of salt bridges	No. of disulphide bonds	No. of hydrogen bonds	No. of non-bonded contacts
AH	13:19	861:823	1	-	7	121
AL	1:2	141:118	-	-	-	8
HL	35:38	1813:1840	2	-	11	214

G protein coupled receptor

Chains	No. of interface residues	Interface area (Å²)	No. of salt bridges	No. of disulphide bonds	No. of hydrogen bonds	No. of non-bonded contacts
AH	15:15	700:801	-	-	6	87
AL	8:14	544:501	-	-	8	68
HL	39:42	1960:1874	2	-	24	321

Metalloproteinase inhibitor 1

Chains	No. of interface residues	Interface area (Å²)	No. of salt bridges	No. of disulphide bonds	No. of hydrogen bonds	No. of non-bonded contacts
AH	15:12	579:668	2	-	4	89
AL	8:8	396:403	-	-	5	31
HL	39:42	1961:1920	2	-	22	318

Lipoprotein lipase

Chains	No. of interface residues	Interface area (Å²)	No. of salt bridges	No. of disulphide bonds	No. of hydrogen bonds	No. of non-bonded contacts
AH	23:20	919:978	1	-	16	148
AL	5:4	182:202	-	-	5	17
HL	40:43	1969:1920	2	-	24	322

Nidogen-1

Chains	No. of interface residues	Interface area (Å²)	No. of salt bridges	No. of disulphide bonds	No. of hydrogen bonds	No. of non-bonded contacts
AH	17:12	592:685	-	-	9	73
AL	10:10	456:477	-	-	8	58
HL	39:43	1992:1949	2	-	24	329

SPARC

Chains	No. of interface residues	Interface area (Å²)	No. of salt bridges	No. of disulphide bonds	No. of hydrogen bonds	No. of non-bonded contacts
AH	8:9	508:479	1	-	4	78
AL	17:17	937:946	3	-	13	119
HL	40:39	1932:1906	2	-	22	313

Sulfated glycoprotein 1

Chains	No. of interface residues	Interface area (Å²)	No. of salt bridges	No. of disulphide bonds	No. of hydrogen bonds	No. of non-bonded contacts
AH	15:14	795:723	1	-	10	94
AL	40:41	1954:1916	2	-	23	326

Figure 62: Interaction Profile summary for each protein showing the number of interactions between different protein-protein chains, out of Salt Bridges, Disulphide Bonds, Hydrogen Bonds, Non-Bonded contacts i.e., Vander wall forces etc.

Interaction Profile summary for each protein showing the number of interactions between different protein-protein chains, out of **Salt Bridges** (a link between electrically charged acidic and basic groups, especially on different parts of a large

molecule such as a protein), **Disulphide Bonds** (is a covalent bond derived from two thiol groups), **Hydrogen Bonds** (a weak bond between two molecules resulting from an electrostatic attraction between a proton in one molecule and an electronegative atom in the other), Non-Bonded contacts i.e., Vander wall forces etc. which are possibly the reason for the HCP's eluting with the mAb purification. It was found that **Cathepsin D** has a total of **1 salt bridge** and **16 hydrogen bonds**, **Clustrin** has **4 salt bridge** and **20 hydrogen bonds**, **Galectin-3-binding protein** has **3 salt bridge** and **18 hydrogen bonds**, **G protein coupled receptor** has **2 salt bridge** and **37 hydrogen bonds**, **Insulin-like growth factor 2 RNA-binding protein** has **4 salt bridge** and **36 hydrogen bonds**, **Lipoprotein lipase** has **3 salt bridge** and **45 hydrogen bonds**, **SPARC** has **6 salt bridge** and **39 hydrogen bonds**, **Metalloproteinase inhibitor 1** has **4 salt bridge** and **31 hydrogen bonds**, **Nidogen-1** has **2 salt bridge** and **41 hydrogen bonds** and **Sulfated glycoprotein 1** has **3 salt bridge** and **33 hydrogen bonds.** Thus, these bonds indicate the possible reasons for the co elution of HCPs with the mAb during its purification.

10) **Design the wash process accordingly:** Based on these interaction profiles there is a need to design a modified wash process accordingly to break the salt bridges and hydrogen bonds between the HCP's and monoclonal antibodies. Without compromising the quality of the monoclonal antibody. As till now elution buffer is used multiple times for the washing of the harvest which has not proven to be that effective.

CONCLUSION

Varied uses of monoclonal antibody in various treatment due to their effectivity and least side effects make it necessary to obtain a risk-free product, but the impurities like host cell proteins make it difficult to obtain such results due to the possible side effects of HCPs such as autoimmune responses, allergies and degradation of drug due to HCPs acting as proteases. As these are used for treatment of deadly diseases with very less medicines effective in trials, there are limited options and people need to rely on these.

Thus, the study of their interaction profile becomes necessary to get a clear picture of their interactions which remain intact after washing with elution buffer multiple times. Results have depicted that they mostly have salt bridges and hydrogen bonds leading to HCPs bonding with the monoclonal antibodies, salt bridges and hydrogen bonds also contribute in conformational structure of proteins, thus, it is necessary to verify these interactions using wet lab methods, also that while designing the wash process it should be kept in mind that the structure of mAb should stay unchanged in its medicinal properties. Thus, it's a long process of testing lots of wash buffers to optimize the best buffer in order to improve the effectiveness of these medicines.

FUTURE PROSPECTS

Study of these interaction profile gave an insight to designing the wash process for the removal of these HCPs, but it needs extensive repetitive wet lab testing to optimize best wash process for the removal of HCPs without affecting the original mAb configuration. Further study of these interactions could be done by studying the conformation of these proteins using crystallization. Later, once the results are positive further work could be carried out on designing the wash process by using different buffers and using mass spectrometry analysis to verify that these HCPs are removed.

Once the wash process is optimized for one monoclonal antibody similar study could be replicated for other monoclonal antibodies available for treatment to get the best results.

It has an extensive application in pharmaceutical industry, once its verified in wet lab and backed up with experimental data.

REFERENCES

1. Ansar, W., & Ghosh, S. (n.d.). Monoclonal Antibodies : a tool in clinical research. https://doi.org/10.4137/IJCM.S11968

2. Cheng, J., Li, J., Wang, Z., Eickholt, J., & Deng, X. (2012). The MULTICOM toolbox for protein structure prediction.

3. Colovos, C., & Yeates, T. (1993). Verification of protein structures : Patterns of nonbonded atomic interactions, 1511–1519.

4. Conchillo-Solé, O., de Groot, N. S., Avilés, F. X., Vendrell, J., Daura, X., & Ventura, S. (2007). AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics*, *8*(February 2007), 65. https://doi.org/10.1186/1471-2105-8-65

5. Jayaram, B., Dhingra, P., Mishra, A., Kaushik, R., Mukherjee, G., & Singh, A. (2014). Bhageerath -H : A homology / ab initio hybrid server for predicting tertiary structures of monomeric soluble proteins, *15*(Suppl 16), 1–12.

6. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling , prediction and analysis. *Nature Protocols*, *10*(6), 845–858. https://doi.org/10.1038/nprot.2015-053

7. Kim, D. E., Chivian, D., & Baker, D. (2004). Protein structure prediction and analysis using the Robetta server, *32*, 526–531. https://doi.org/10.1093/nar/gkh468

8. Kozakov, D., Hall, D. R., Xia, B., Porter, K. A., Padhorny, D., Yueh, C., … Vajda, S. (2017). The ClusPro web server for protein–protein docking. *Nature Protocols*, *12*(2), 255–278. https://doi.org/10.1038/nprot.2016.169

9. Laskowski, R. A. (2001). PDBsum : summaries and analyses of PDB structures, *29*(1), 221–222.

10. Leiss, M., Pester, O., & Aschner, M. (2015). Pharmaceutical Getting CHO host cell protein analysis up to speed, *3*, 13–23.

11. Levy, N. E., Valente, K. N., Choe, L. H., Lee, K. H., & Lenhoff, A. M. (2014). Identification and characterization of host cell protein product-associated impurities in monoclonal antibody bioprocessing. *Biotechnology and Bioengineering*, *111*(5), 904–912. https://doi.org/10.1002/bit.25158

12. Levy, N. E., Valente, K. N., Lee, K. H., & Lenhoff, A. M. (2016). Host Cell Protein Impurities in Chromatographic Polishing Steps for Monoclonal Antibody Purification, *113*(6), 1260–1272. https://doi.org/10.1002/bit.25882

13. Li, F., Vijayasankaran, N., Shen, A., Kiss, R., & Amanullah, A. (2010). Cell culture processes for monoclonal antibody production. *mAbs*, *2*(5), 466–479. https://doi.org/10.4161/mabs.2.5.12720

14. Schenauer, M. R., Flynn, G. C., & Goetze, A. M. (2012). Identification and quantification of host cell protein impurities in biotherapeutics using mass spectrometry. *Analytical Biochemistry*, *428*(2), 150–157.

https://doi.org/10.1016/j.ab.2012.05.018

15. Singh, A., Kaushik, R., Mishra, A., Shanker, A., & Jayaram, B. (2016). Biochimica et Biophysica Acta ProTSAV : A protein tertiary structure analysis and validation server. *BBA - Proteins and Proteomics*, *1864*(1), 11–19. https://doi.org/10.1016/j.bbapap.2015.10.004

16. Thornton, J. (1993). PROCHECK : A program to check the stereochemical quality of protein structures, (March 2014). https://doi.org/10.1107/S0021889892009944

17. Valente, K. N., Lenhoff, A. M., & Lee, K. H. (2015). Expression of Difficult-to-Remove Host Cell Protein Impurities During Extended Chinese Hamster Ovary Cell Culture and Their Impact on Continuous Bioprocessing, *112*(6), 1232–1242. https://doi.org/10.1002/bit.25515

18. Valente, K. N., Schaefer, A. K., Kempton, H. R., Lenhoff, A. M., & Lee, K. H. (2014). Recovery of Chinese hamster ovary host cell proteins for proteomic analysis. *Biotechnology Journal*, *9*(1), 87–99. https://doi.org/10.1002/biot.201300190

19. Wlaschin, K. F., & Yap, M. G. S. (1987). Recombinant Protein Therapeutics from CHO Cells — 20 Years and Counting, 40–47.

20. Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite : protein structure and function prediction. *Nature Publishing Group*, *12*(1), 7–8. https://doi.org/10.1038/nmeth.3213

21. Zola, H. Monoclonal antibodies. In: Encyclopedia of Life Sciences. John Wiley & Sons: Chichester, UK; 2010:1–9.

22. Tyagi S, Sharma PK, Kumar N, Visht S. Hybridoma technique in pharmaceutical science. International Journal of Pharm Tech Research. 2011;3(1):459–463.

23. Edwards PA. Some properties and applications of monoclonal antibodies. Biochem J. 1981;200(1):1–10.

24. Köhler G, Milstein C. Continuous cultures of fused cells secreting antibody of predefined specificity. Nature. 1975;256(5517):495–457

25. B. Leader, Q.J. Baca, D.E. Golan, Protein therapeutics: a summary and pharmacological classification, Nat. Rev. Drug Discov. 7 (2008) 21–39.

26. J.M. Reichert, Antibody-based therapeutics to watch in 2011, MAbs 3 (2011) 76–99.

27. K. Maggon, Monoclonal antibody ‘‘gold rush’’, Curr. Med. Chem. 14 (2007) 1978–1987.

28. E.W. Leser, J.A. Asenjo, Rational design of purification processes for recombinant proteins, J. Chromatogr. 584 (1992) 43–57.

29. A.A. Shukla, J. Thommes, Recent advances in large-scale production of monoclonal antibodies and related proteins, Trends Biotechnol. 28 (2010) 253–261.

30. P. Gagnon, Polishing methods for monoclonal IgG purification, in: A. Shukla, S.

31. Gadam, M. Etzel (Eds.), Process Scale Bioseparations for the Biopharmaceutical Industry, Taylor & Francis/CRC, Boca Raton, FL, 2007, pp. 491–505.

32. Ma J, Wang S, Zhao F, Xu J. Protein threading using context-specific alignment potential. *Bioinformatics (Proceedings of ISMB 2013)*. 2013 Jul 1;29(13):i257-65. doi: 10.1093/bioinformatics/btt210.

33. Jianzhu Ma, Jian Peng, Sheng Wang and Jinbo Xu. A conditional neural fields model for protein threading. *Bioinformatics (Proceedings of ISMB 2012)*, 2012.

34. Jian Peng and Jinbo Xu. RaptorX: exploiting structure information for protein alignment by statistical inference. *PROTEINS*, 2011.

35. Jian Peng and Jinbo Xu. A multiple-template approach to protein threading. *PROTEINS*, 2011.

36. Jian Peng and Jinbo Xu. Low-homology protein threading. *Bioinformatics (Proceedings of ISMB 2010)*, 2010.

37. Jian Peng and Jinbo Xu. Boosting protein threading accuracy. *In the Proceedings of the 13th International Conference on Research in Computational Molecular Biology (RECOMB), Lecture Notes in Computer Science*.

38. Wang Z, Zhao F, Peng J, Xu J. Protein 8-class secondary structure prediction using conditional neural field.

39. Zhiyong Wang and Jinbo Xu. Predicting protein contact map using evolutionary and physical constraints by integer programming. *Bioinformatics (Proceedings of ISMB 2013)*, 2013.

40. Jianzhu Ma, Sheng Wang, Zhiyong Wang and Jinbo Xu. Protein Contact Prediction by Integrating Joint Evolutionary Coupling Analysis and Supervised Learning. *RECOMB 2015, Lecture Notes in Computer Science*, Volume 9029, 2015, pp 218-221.

41. Feng Zhao, Jian Peng and Jinbo Xu. Fragment-free approach to protein folding using conditional neural fields. *Bioinformatics (Proceedings of ISMB 2010)*, 2010.

42. Feng Zhao, Jian Peng, Joe DeBartolo, Karl F. Freed, Tobin R. Sosnick and Jinbo Xu. A probabilistic and continuous model of protein conformational space for template-free modeling. *Journal of Computational Biology*, 2010.

43. Kozakov D, Hall DR, Xia B, Porter KA, Padhorny D, Yueh C, Beglov D, Vajda S. The ClusPro web server for protein-protein docking. *Nature Protocols.*2017 Feb;12(2):255-278.

44. Kozakov D, Beglov D, Bohnuud T, Mottarella S, Xia B, Hall DR, Vajda, S. How good is automated protein docking? *Proteins: Structure, Function, and Bioinformatics*, 2013 Aug.

45. Kozakov D, Brenke R, Comeau SR, Vajda S. PIPER: An FFT-based protein docking program with pairwise potentials. *Proteins*. 2006 Aug 24.

46. Comeau SR, Gatchell DW, Vajda S, Camacho CJ. ClusPro: an automated docking and discrimination method for the prediction of protein complexes.*Bioinformatics*. 2004 Jan 1.

47. Comeau SR, Gatchell DW, Vajda S, Camacho CJ. ClusPro: a fully automated algorithm for protein-protein docking *Nucleic Acids Research*. 2004 Jul 1.

48. Morris AL, MacArthur MW, Hutchinson EG, Thornton JM (1992). Stereochemical quality of protein structure coordinates. Proteins, 12, 345-364.

49. Kremkow BG, Baik JY, MacDonald ML, and Lee KH. "CHOgenome.org 2.0: Genome resources and website updates." Biotechnol J (2015) 10, 931-938.

50. Hammond S, Kaplarevic M, Borth N, Betenbaugh MJ, and Lee KH. "Chinese Hamster Genome Database: An Online Resource for the CHO Community at www.CHOgenome.org."Biotechnol Bioeng (2012) 109, 1353-1356.

Appendix

Table of Figures