Acta Chimica and Pharmaceutica Indica
ISSN: 2277-288X

Acta Chimica and Pharmaceutica Indica

A modernised fast tool for computational structure prediction

Subin Mary Zachariah*

Department of Pharmaceutical Chemistry and Analysis, Amrita School of Pharmacy, Kerala, India

Abstract

Drug development and drug discovery are time-consuming and involve many procedures, but several methods introduced modifies the previous making it more easy and simplified. Also, these methods produce more accuracy, and models are made in a short time and take less expenditure in terms of both processing and verifying. For this, appropriate protein structures are selected and their interactions are studied, and based on it new models are made. This involves structural similarity which is one of the key features and their binding on to specific target sites which should also be considered. Based on this criteria model or mannequins are made and their action on specific sites is studied. Methods such as X-ray crystallography and NMR are often used these are not accurate and models formed more complex and difficult to study.

The basic strategy behind the model building is the formation of macromolecule -ligand complexes and protein-ligand interactions which are very important and gives detailed information for new drug development. Due to the absence of experimental data, model building on this concept has been difficult and time-consuming therefore by using framed 3D structure of a homologous organic compound is one of the only reliable techniques to induce the structural information in today's arena. Information of the 3D structures of proteins provides valuable insights into the molecular basis of their functions. The recent advances in similarity modeling, mainly in investigation and sequencing with model structures, distant homologs, modeling of loops and side chains additions contributed to the consistent prediction of organic compound structure, that wasn't realizable even a few years past. But now the method of similarity modeling made this easier. Error occurred during study was checked and validated, again done in an accurate manner.

This review, therefore gives rough information on how similarity modeling is done, various steps involved and validation. Information about newer methods and advances are also mentioned. And finally concluded with applications and their importance in drug development. Henceforth homology modeling has been considered as one of the most used and advanced methodologies for the prediction of proteins, their alignment and creation of new models for drug development.

Keywords:

Drug discovery, GPCRs, Homology modeling, Ligand design, Loop structure prediction, Model validation, Sequence alignment

Introduction

Major problems faced for drug design and drug development is the prediction of three dimensional structures of proteins having its own sequences which has highest accuracy and is comparable rapidly, in cost effective manner. As most of drugs have these structural protein molecules as their target. Due to the need of high specificity for drug target protein and its use for drug discovery and structure based drug design is highly limited. Henceforth, protein target identification and its verification are one of the major aspects for drug discovery projects.

Experimental methods have been difficult to study in elaborate about the protein structure, its interaction and molecular docking which is often used. This includes NMR and X-ray crystallography of protein. NMR -nuclear magnetic resonance spectroscopy which deals with molecular physics, gives the information regarding the structure of proteins, nucleic acids, and their related complexes using simplified methods of molecular dynamics but the structures formed are often distorted. Another method used is X-RAY crystallography which includes an examination of protein structure in crystallized form [1]. These methods are basic and give the whole 3D structure that is later required for new protein discovery or new drug development.

Citation: Subin Mary Zachariah, A modernised fast tool for computational structure prediction, September 09, 2021.

But methods used have found to be more time-consuming in terms of obtaining sufficient materials which include cloning, expression, crystallization, and purification of protein and is relatively more expensive. Also, this method has several demerits and is not successful for larger proteins. Having associated with these difficulties in coming modern world protein X-ray crystallography and NMR method of analysis would not be preferred. This brings our keen interest in protein modeling using computer databases. A database usually consists of a template sequence that is of known 3D structure present in the protein data bank, through which we can produce target protein model by its alignment with the query sequence or an unknown 3D structure of a protein. This is the basic principle behind homology modeling.

Model quality depends on the accuracy of sequence alignment and structural quality of the template. Identification of template sequence should be done explicitly. This technique of modeling through homology or by comparison with known structure (template), protein structure prediction helps in easy identification of target proteins. However, the quality of constructed protein model is linked with percentage similarity between target and template. If the model is having 30%-50% of accuracy, then it is apt for drug design. This method therefore solves the problem of prediction of 3D structural proteins from amino acid of its own.

Literarture Review

Through this method we can calculate the protein free energy, by using which we can find the global minimum. Protein free energy is the ability of protein molecule to fold into their highly structured functional state. It can be calculated by Gibbs free energy difference, which gives protein stability.

Basic aim of homology is to obtain greater knowledge about the protein structures, solving the unsolved protein sequences in order to know its functions and detecting errors. This contributes to prediction of protein structure. It fastens the details through visualization techniques and differential properties of proteins can be discovered. It is a tool to modeling ligand structural models, mutagenesis experiments and loop structure prediction. This article gives knowledge of understanding the role and reliability of homology modeling in drug discovery and development process.

Template fold recognition structure of protein of unknown sequences is determined based on its alignment with known sequence of proteins. On evolution structures have become more stable and show very little changes with that of associated sequences, such that sequences, which are similar would adopt themselves into identical structures and those which are distantly related would fold into similar structures. The unknown sequence is compared with the known sequences for alignment, present in protein data bank that can be any of the programs specified. Depending on percentage of sequence similarity known sequence should completely match with that of unknown sequence and the known sequence taken is the template strand.

The server used commonly BLAST21 (Basic Local alignment search Tool) is one of the popular servers and a search on it gives a list of known sequence of proteins which is then compared with that of the query and alignment is made. This known sequence is called template and the server finds template only if the sequence identity is below 30% ie; alignment made should be 47%-50%. Errors on alignment should be corrected and can be the cause for the deviation in comparative modeling even when correct template is chosen. Numerous steps are involved in template recognition and sequence alignment and these steps are facilitated by structure databases and database scanning software's such as SAM, HMMER, CLUSTAL W or CLUSTAL. The protein comparison takes place based on two classes which are useful methods for identification and fold recognition.

I. Class include a comparison of target sequence with each of sequences present in database independently using a pair wise sequence-sequence identity. BLAST is frequently used programs in this class; this can also include FASTA and CDART. But this class of method can give only half of evolutionary relationship having a range of sequence identity of about 20%-30%.
II. Class -Using multiple sequence information a comparison with the sequence is done by using profile analysis, profile-profile comparisons and Hidden Markov Models and Intermediate sequence search. Popularly used programs include PSI-BLAST.
III. Class- Threading or 3D template matching method. It mainly includes a pair wise comparison of protein sequence with protein of known structure. Even if target sequence is not known it adopts any one of known 3D folds that has been already optimized through scoring and it threads along with 3D folds i.e.; independently for each sequence-structure pair. And commonly used program is THREADER.

A comparison of query sequence with sequences of known structures is done using matrices residue exchange matrix containing 20 amino acids is drawn out of which two of them would align and alignment is based on physic -chemical properties and each has its own score. Residues highly conserved generally get the highest score. The axes of this matrix consist of two sequences that have to be aligned. During the alignment process, a best path is found through this matrix. This makes sure that no residue is used twice; one must consider taking at least one step to the right and one step down. Non-identical residues are indicated through gaps in a row. The optimum path corresponding to the alignment on the right side is shown in grey. Residues with properties which are similar are marked with a '#' while the dashed line is marked as an alternative alignment. Sequences which are highly similar will have higher score and is suitable for alignment. In practice, template structures can be easily obtained by putting the query sequence to one of the BLAST servers on the web and is searched using the PDB as the database [2]. The template structure having the highest sequence identity is made to run across the query sequence and percentage is identified.

Use of parameters such as the present cofactors, multimedia complexes or conformational states (active or inactive) or any other

molecules will give a greater impact on model building. Also advances in programming and multiple software's, made it easier to select appropriate templates and build multiple models which has helped the research analyst to choose best model for further study. It is also easier now to combine multiple templates into one structure that is used for modeling. The online Swiss model and servers better use this approach.
Alignment correction sequence identity plays major role to obtain an accurate alignment. Alignments having sequence identity of range less than 30% is most valid but have greater chances of errors. Pair wise sequences aligned with this sequence identity have only very few residues aligned correctly remaining 20% of sequence residues is aligned incorrectly. In some cases the pair of related proteins has no correct aligned positions when aligned by sequence based alignment methods.38Alignment accuracy at the safe zone is very crucial for wide applications, mainly in comparative protein structure prediction. To get the accurate comparative model, it is important to get at least one correctly aligned target with that of the template. An incorrect alignment gives an inaccurate model .Therefore alignments have to be optimized specifically for comparative protein structure prediction. This is done by 13 profile-profile alignment protocol ie; it converts multiple sequence alignment into profile or a matrix and a comparison is done between the two profiles. Alignment sets are then tested and alignment accuracy is measured. PSI-BLAST method is used for sequence profile alignment with most significant e-value.
Calculation of sequence profile from multiple sequence alignment is done by 3 methods:

1. Sequence weighting
2. Sequence profile
3. Profile -Profile substitution scores

These are different algorithm method used to calculate the accuracy of alignment. Alignments were randomly divided into training and testing sets the training set of alignment optimizes gap initiation and gap extension penalties, while the testing set was used to assess the performance of all examined alignment methods. The accuracy of the alignment was measured by relying on aligned native structures derived from PDB. Alignment methods were assessed by the percentage of alignment with the structure overlap if it's greater than 30%. Then the structure pairs with maximum overlap. The results of these calculated profile-profile alignments were then observed. In some cases when two sequences are difficult to align at exact positions as the percentage of sequence identity is comparatively low (Multiple sequence alignment). With the two sequence a third sequence is used which easily align with the two sequence and resolves the issue.

As the alignment is now ready for modal building, the generation of backbone is the initial step for this. Experimentally determined protein structures shows wide range of errors such as poor electron density in X-ray diffraction, due to this a large number structures present in PDB ie; about 50000 structures shows a wide range of errors which is double of it. Due to this current X-ray and NMR structures are now re-refined. Re-refinement before modeling is now more appropriate. If there are two templates then both are combined as done in multiple template modeling. And the template chosen must be of least errors. Modal building starts through backbone generation –This involves assembling of rigid bodies from the aligned protein structures. Followed by natural dissection of protein structures into conserved core bodies, variable loops that connect them and side chains that decorate the backbone.

Modeling of a protein has much application in the prediction of structure and designing of molecules. This protein side chain makes a major step in the modeling method. Protein side chain usually exists in low energy roamers. One of the important methods in homology is side chain modeling. This method is usually done by placing the side chain to the coordinates at the backbone which is obtained from the parent chain. This type of side chain is seen in chains which have low energy roamers. In the side chain finding, the method roamer is determined using the sequence of protein structure and the coordinates of the backbone. A group of roamers is used to make side chains having 5 -6 conformations in each. In the chain, the lowest energy combination is identified and added to the backbone. The higher the accuracy more appropriate for modeling the protein structure.

Errors inside the mannequin are substantially common and most attention is required closer to refinement and validation. Errors in mannequin are commonly calculable by way of superposition of mannequin onto native structure. Mathematics and calculations important between matched structures for the model as it scores to point out smart structural similarity and for development of a contrast operate which is capable of discriminating true models. Applied mathematics and high-quality strength features are supported to the located homes of amino acids in well-known structures. An unfold of applied mathematics standards are derived for assorted properties like distributions of polar and polar residues within or backyard of macromolecule, thereby knowing the misfiled models. Certain associations will locate the native errors and entire misfields; packing rules are enforced for shape evaluation. A model which is alleged will be now valid solely if some distortions present in atomic characters are corrected. The Ramachandran plot is possibly the major effective determinant of the difference in protein. Facet chain torsion angles has indispensable atomic number, usually altered in the course of the modeling method.

Conformational free energy distinguishes the native shape of a macromolecule from other in degree. One of the benefits of such physically derived functions is that they support well-defined physical interactions; consequently it is difficult to attain perception from their performance [3]. Additionally, ab-initio strategies confirmed success in recent CASP. One in all the principal drawbacks of physical chemical description of the folding free energy of a macromolecule is that the methods typically come at a large procedure and more expenditure. Quick association models like the generalized born and a spread of simplified assessment schemes ought to have an effect and is very beneficial.

Variety of freely available applications may affirm similarity models, amongst them WHAT_CHECK solves normally crystallographic

problems. The validation programs are usually of 2 types: (1) preliminary class (e.g. PROCHECK and WHATIF) exams for correct macromolecule stereochemistry, like symmetry checks, pure arithmetic exams (chirality, bond lengths, bond angles, torsion angles models). Solvation and structural packing is done and the 2d class (e.g.VERIFY3D and PROSAII) checks the fitness of sequence to shape and assigns a rating for each residue becoming its contemporary surroundings. GRASP2 is new mannequin evaluation package developed via Honig. As an example, gaps and insertions may also be mapped to the structures to verify that they create field geometrically. It's suggested that, guide scrutiny should be blended with current applications to resolve other issues inside the model.

One of the main methods used for prediction of a super molecule is ab initio method .This methodology is generally most admired for structure prediction as soon as there is no or terribly low quantity of similarity for the supermolecule.53 It's one of the hardest established strategy having a random conformation. The ab-initio methodology depends on the natural philosophy hypothesis projected via Anfinsen, as per that the native shape corresponds to the free energy minimum (gibbs free) under a given set of conditions. There are many ab-initio structure prediction procedures offered like ROSETTA, TOUCHSTONE-II, and also the most commonly famous I-Taser. These methods supported the Monte-Carlo algorithmic rule. It's been known that I-Tasser outperforms the ROSETTA and TOUCHSTONE-II tactics with a way lower mainframe value [4]. The ab-initio modeling is normally termed as de-novo modeling, physics-based modelling, or free modelling. The quintessential protocol accompanied by using the ab-initio approach of the super molecule structure prediction is evolved with the first natural compound sequence that is probe with larger number of conformations resulting in the prediction of native folds. When the folds are diagnosed and foretold, the model assessment is performed to verify the modest of the predicted structure. ROSETTA and I-Tasser comply with the accelerated methodology for ab-initio prediction of a super molecule.

ROSETTA prediction starts with the identification of little fragments (timers and nanomers) from the structure databases that have consistency with native sequence preferences. After that, all the fragments are assembled into models with world residences accompanied by means of the assessment employing a marking function from decoy population. The protocol followed by using the I-Tasser includes threading collectively with the ab-initio techniques. I-Taser application is predicated on the secondary-structure accelerated Profile-Profile threading Alignment (PPA) and additionally reiterated followed by implementation of the Threading Assembly Refinement (TASSER) program.

Structure-based drug design technique in most refined and produce accurate models having characteristic amino acid sequence and similarity protein structures Helps in building a similarity model of the whole super molecule. The aim of homology modeling is to predict a structure from its sequence with an accuracy that's almost like the results obtained through an experiment. Homology modeling is efficient and provides the various methodologies to get models. Modeling studies are mounted through mental image technique, and also differential properties of the proteins may be discovered. The role and dependableness of the homology model building can still grow because of the variety through an experiment determined structures will increase. Homology modeling may be a powerful tool to counsel modeling of ligand-receptor interactions, enzyme-substrate interactions, cause experiments, SAR data, lead optimization and loop structure prediction. Homology modeling powerfully depends on the virtual screening and thriving tying up results. It also gives information about protein binding and accordingly new drug is developed from the existing protein structures in a simplified manner. The use of protein macromolecules from PDB has been more frequent and similarity models are created which are most accurate and least expensive in both processing and producing. Moreover, this modeling also involves mathematical calculations and G-bbs free energy changes for the selection processes, and matching is done. This method also involves the re-correction of complicated large protein molecules in the form of loops and model validation is done by using macromolecule mannequin and verified accordingly. The method is more simplified and modifications are noted [5]. These recent advances ought to facilitate to enhance our data of understanding the role of homology modeling in drug discovery method.

References

1. Cavasotto CN, Phatak SS. Homology modeling in drug discovery: Current trends and applications. Drug Discov. 2009;14(13-14):676-83.
2. Chandonia JM, Brenner SE. Implications of structural genomics target selection strategies: Pfam 5000, whole genome, and random approaches. Proteins. 2005;58(1):166-79.
3. Vitkup D, Melamud E, Moult J, et al. Completeness in structural genomics. Nat Struct Biol. 2001;8(6):559-66.
4. Floudas CA, Fung HK, McAllister SR, et al. Advances in protein structure prediction and de novo protein design: A review. Chem Eng Sci. 2006;61:966–88.
5. Equeenuddin Sk, Tripathy S, Sahoo PK, et al. Hydrogeochemical characteristics of acid mine drainage and water pollution at Makum Coalfield, India. J Geochem Expl. 2010;105:75-82.