View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by University of Queensland eSpace

Dr. Megan L. O'Mara and Dr. Evelyne Deplazes

Final Author Version

Polypeptide and Protein Modeling for Drug Design

Encyclopedia of Computational Neuroscience

Synonyms

Computational modeling, structure-aided drug design, computational structural biology

Definition

Modeling of polypeptides and proteins, often referred to as biomolecular or molecular modeling, encompasses the use of theoretical models and computational methods to model the structure and dynamics of molecules of biological interest such as peptides, proteins and small organic molecules (ligands).

The aims of protein and peptide modeling for drug design include: (i) modeling the three-dimensional structure of proteins of current or potential drug targets; (ii) identifying and characterizing the structural dynamics associated with the function of a particular protein or peptide; and (iii) predicting the structure and molecular interactions of protein-ligand complexes. This knowledge underpins structure-based and rational drug design. It can aid in the design and optimization of drug molecules by shedding light on their mode of action, specificity and selectivity.

Detailed Description

Background

The number of protein structures solved experimentally and deposited in the Protein Data Bank (PDB) has risen exponentially since its inception the 1970's (RSCB Protein Data Bank). This has been accompanied by a rapid growth in computational power and continuous improvements in modeling algorithms, greatly enhancing the complexity, size and diversity of the proteins that can be simulated. Together, the experimental and computational breakthroughs initiated in the 1970's have contributed enormously to our knowledge of protein structure and structural conservation within protein superfamilies. In 2013, the contribution of computational modeling to our understanding of protein structure and biochemical function was recognized by the award of the Nobel Prize in Chemistry to three of the founders of the field, Martin Karplus, Michael Levitt and Arieh Warshel (The Nobel Prize in Chemistry 2013 - Press Release 2013). Computational modeling techniques lie at the core of fields such as protein folding and dynamics, structure-based drug design and computer-aided drug design.

The majority of drugs targeting the central nervous system modulate the action of one or more neurotransmitters involved in synaptic transmission. Thus the most common drug targets fall into the category of cell surface receptors, such as metabotropic and G-coupled protein receptors (GPCRs), transporters and ion channels (Cross and Yocca 2012; Hefti 2004; Squire et al. 2008). In fact, drugs targeting GPCRs account for over 50% of all clinical drugs (Schushan and Ben-Tal 2010). Despite their clinical significance, solving the experimental structure of transmembrane proteins remains challenging. In 2013, the PDB contained >94500 structures, of which only 1954 are transmembrane proteins (RSCB Protein Data Bank ; Kozma et al. 2013). This gap between the availability of structural models and the need for high-resolution structures of membrane proteins is a major obstacle in the design of drugs targeted for a specific protein or protein isoform. In many cases, both the protein dynamics and the precise details of the drug binding site is unknown, making identification of the molecular interactions that govern the specificity of the drug difficult to determine. Furthermore, the existence of multiple subtypes or isoforms of many of these proteins, each with distinct physiologic functions, makes subtype selectivity crucial for drug efficacy. Computational modeling techniques offer a means to understand structure-function relationships of these critical proteins.

The underlying principles behind protein and peptide modeling techniques

Computational modeling is used to address a myriad of question relevant to drug design. Examples include predicting the native structure of a protein from its amino acid sequence or from the structure of homologues; elucidating the relationship between protein structure and biological function; understanding how protein motion is associated with drug binding; determining the molecular interactions between a drug and its target protein that govern drug selectivity and potency; and understanding how drug binding modulates protein function. The most suitable modeling approach for a particular problem depends on the level of structural and biochemical information available and the question to be addressed. Figure 1 gives an overview of commonly used techniques that will be discussed here.

Simulation techniques for modeling of proteins, peptides and drugs can be classified into three broad categories based on the underlying theory or principle: knowledge or rule based methods; potential energy methods; and quantum mechanical methods (Goodfellow 2008). More recent approaches combine two or more of these categories to improve the quality of the models obtained (Goodfellow 2008). For example, computational approaches that investigate protein dynamics, or protein–ligand interactions (such as molecular dynamics and docking techniques) are primarily based on potential energy methods, while structure prediction methods often combine an initial knowledge-based method with a potential energy method to further refine the model. Methods based on a purely quantum mechanical approach are not feasible for molecules of more than 50 atoms due to current computational limitations. As most proteins and peptides are larger than this, quantum mechanical approaches will not be discussed here.

Knowledge based methods

Knowledge based methods for protein and peptide modeling use comparative statistical analysis to predict the structure of the amino acid sequence of the target protein from (i) the sequence of homologous proteins of known structure; or (ii) a database of protein structures (Rangwala 2010). The accuracy of this approach relies on the current knowledge of protein structures and evolutionary relationships.

Potential energy methods

Potential energy (PE) methods, are based on (i) a description of the fundamental parts of the system (usually atoms), (ii) a mathematical description of the interactions between the atoms; (iii) a starting configuration such as the coordinates of each atom in the system; and (iv) an algorithm to carry out the desired computational steps by evolving the system in time and space (Goodfellow 2008; Jensen 2007)

Most PE methods used for modeling polypeptides and proteins utilize the so-called molecular mechanics (MM) approach in which a system is described as a collection of spheres (representing one or more atoms), connected by bonds in the form of harmonic springs. Both the bonded and non-bonded interactions between all pairs of atoms are described by simple mathematical expressions, referred to as the forcefield, the general form of which is described in Eqn. 1:

$$E_{total} = E_{bond} + E_{angle} + E_{dihedral} + E_{electrostatic} + E_{van\ der\ Waal} \qquad (1)$$

where E is the potential energy. The underlying principle of MM assumes that the potential energy of the molecular system can be described as a function of the atom's position in space. The forcefield allows the change of energy in the system to be calculated as the atomic positions change. The conformation of a protein or peptide is optimized with respect to its energy, as the assumption of the potential energy concept is that the minimization of the potential energy will give the optimal geometry of the protein or ligand. MM becomes particularly powerful when combined with algorithms that propagate the system in time or space (Jensen 2007; Náray-Szabó et al. 2012; Nowak 2012). These algorithms either propagate possible conformations of the protein to sample the "conformational space"; or examine how the system evolves over time. Energy minimization and Monte Carlo algorithms are used to propagate the system in conformational space. Energy minimization algorithms (also referred to as geometry or energy optimization) calculate the gradient of the potential energy and evolve the protein in the direction of the minimum potential energy. Monte Carlo algorithms (named after the Monte Carlo casino) rely on repeated random sampling of an initial amino acid conformation to find the statistically most probable end state for the given initial conditions. These algorithms are particularly important for finding stable, minimum energy conformations of a protein, for example, the local and global energy minima in protein folding or ligand binding pathways. They cannot give information about the timescales of these processes.

Langevin (or stochastic) dynamics and molecular dynamics use algorithms to examine the time-evolution of the system but differ in the underlying equation of motion. Molecular dynamics is based on Newton's equation of motion:

$$m_i a_i = F_i(r) \qquad (2)$$

where m_i, a_i *and* $F_i(r)$ are the mass, acceleration and force on the *i*th atom. Stochastic dynamics is based on the Langevin equation:

$$m_i a_i = F_i(r) - m_i \gamma_i(t) + \eta_i(t) \qquad (3).$$

Here the additional terms, $m_i\gamma_i(t)$ and $\eta_i(t)$ are the random force and frictional force on the *i*th atom that arise from molecular collisions due to the thermal motion of the atoms. By incorporating these forces, Langevin dynamics model the effects of water and other solvents without including them explicitly in the simulations. In contrast to algorithms that propagate the protein system in conformational space, algorithms that give the time evolution of the system provide information about the dynamics of the system and the timescale at which these occur. One of the major limitations of these algorithms is they often evolve to a local energy minima which is stable over the timescale (ns to μs) of the simulation. Because of this limitation, time-evolution algorithms cannot currently sample the entire range of conformations that a protein system will adopt in a physiological environment.

Protein structure prediction

Homology modeling

Proteins derived from a common ancestor are homologues and display a conserved amino acid sequence and three-dimensional structure. The extent of the sequence and structural conservation depends on the evolutionary distance between the homologues. Homology or comparative modeling exploits these similarities to construct a structural model of the target protein based on the experimental structure of homologue, referred to as the modeling template. Conservation of protein folds between homologues guides the modeling, which follows four steps: fold recognition and template selection, target-template alignment, model building, and model assessment (Zvelebil and Baum 2008).

The reliability of the homology model is critically dependent on the accuracy of the sequence alignment between the homologues and the resolution of the experimental structure of the homologue used as the modeling template. Single residue shifts in the alignment of the sequences can result in register shifts across large regions of the protein, seriously affecting the predictions of binding site residues and ligand coordination (Náray-Szabó et al. 2012; Zvelebil and Baum 2008). This is particularly relevant for the structure of transmembrane proteins such as channels, transporters and receptors. The general role of these proteins is to selectively allow the passage of a hydrophilic or charged moiety across a hydrophobic lipid bilayer. To facilitate this process, transmembrane proteins possess highly ordered secondary structures in which the exterior of the protein is a hydrophobic shell that shields a substrate-selective hydrophilic core from the lipid environment. Most channels and membrane transport proteins are composed of interacting α-helices. Here a residue shift of a single amino acid in one helix may alter the prediction of the entire interaction interface between two or more helices, and shift the orientation of binding site residues within an α-helix by 100°.

The sequence identity between the target and template proteins gives a limit to the confidence placed on the homology of the two proteins. 90% of the proteins with sequence identities of >30% are structural homologues, while for pairs of proteins with sequence identities of <20%, less than 10% were structural homologues (Rost 1999; Zvelebil and Baum 2008). Aligned proteins with sequence identities ranging from 30% to 20% represent the transition zone from a high-confidence prediction of homology to a low-confidence prediction. This region is referred to as the "twilight zone" for homology modeling (Rost 1999). It should be noted that in transmembrane proteins, substitutions between hydrophobic residues are generally well tolerated, while the amino acid composition of the substrate-selective regions of the protein are tuned for the substrate of interest. For this reason, transmembrane proteins from the same family or superfamily often have sequence identities that lie within this twilight zone, increasing the possibility of alignment errors in homology modeling.

Modeling by fold recognition

In many cases, the three-dimensional structure of a protein is more highly conserved than the amino acid sequence. Threading or fold recognition approaches attempt to find a conserved structure that is compatible with the sequence of the target protein. The target sequence is "threaded" onto each protein in a database of experimentally determined protein structures and a scoring function is used to determine which structure best accommodates the target sequence. A high score indicates the structure is compatible with the target sequence and it is assumed that the target sequence folds in the same manner (Zvelebil and Baum 2008).

Probabilistic modeling of protein interactions

Molecular Docking Techniques

In the case of drug–protein, peptide–protein or protein–protein interactions, computational docking approaches predict an interaction interface between the protein and ligand, which may represent a stable binding mode. Molecular docking essentially involves two steps: (i) generation of a large number of possible configurations of the protein-ligand complex (binding poses); and (ii) scoring of configurations with the aim to distinguish possible protein-ligand configurations from those not physically possible (Xu et al. 2007). A set of example results from a docking simulation is shown in Figure 2.

Docking methods vary significantly in the degree to which they account for the flexibility of the protein and the ligand. The majority of docking algorithms implement a rigid body approach, in which the internal flexibility of both the protein and the ligand is not considered. More sophisticated approaches allow flexibility of the protein side chain or backbone atoms. The choice of method is essentially a trade-off between accuracy and speed. Rigid body methods are fast and robust and can be applied for virtual screening of a large number of potential ligands, albeit at low accuracy. More sophisticated flexible docking methods allow optimization of the protein-ligand complex. These methods are much slower, but provide greater accuracy (Xu et al. 2007). Recent data driven approaches use experimental information such as NMR chemical shift perturbation or mutagenesis data as internal constraints to drive the docking algorithm (Xu et al. 2007).

Scoring functions are used to assess the quality of modeled structures and interactions

Scoring functions are an important concept used for assessing the quality of structural models predicted by either potential energy or knowledge-based methods. Scoring functions are approximate mathematical methods that give an empirical score that represents the mathematical quality of the model. A favorable score should be interpreted in light of the experimental results. It does not necessarily imply that the homology model or docked complex represents the physiological state. Instead, it may simply reflect a well-optimized mathematical solution.

Scoring functions for knowledge-based methods provide a measure of the structural environment of the model. Knowledge-based scoring functions are based on mathematical functions that consider inter-residue distances, such as side chain packing and solvation potential; or solvent accessibility and the fit of the sequence into a given solvent accessible surface (Zvelebil and Baum 2008). Potential energy scoring functions are based on steric fit (or hindrance) of the molecular interactions derived from the van der Waals interactions between adjacent atoms, the energetics of the torsional angles for the covalent interactions, and electrostatic potential energy of the interaction (Baron et al. 2012; Náray-Szabó et al. 2012).

Modeling time-dependent protein dynamics

Molecular Dynamics simulation techniques

The function of most proteins is intrinsically linked to dynamic structural changes. Classical molecular dynamics (MD) simulation techniques have been used to model conformational changes of proteins and peptides on ns to μs timescales. While many biologically important conformational changes, such as protein folding, occur over longer (μs to ms) timescales, MD simulations are able to capture rapidly induced conformational changes, such as those involved in the process of ligand binding (10 -500 ns). Recent advances in algorithm and hardware design, or the implementation of so-called coarse-grained forcefields have extended the possible simulation times to simulate the conformational dynamics of channel gating (500 ns – 5 μs). MD simulations are the most computationally intensive of the techniques discussed here. However, they are currently the only method (computational or experimental) that characterizes the structural dynamics and molecular interactions in atomic detail, while fully accounting for the flexibility of the protein and ligand in a physiologically relevant environment. MD simulations provide an accurate method of studying the molecular interactions involved in ligand binding; the conformational changes induced in the protein or the ligand during the formation of a stable complex; and the correlation of ligand orientation and protein conformations (Shirts 2012).

In general terms, molecular dynamics (MD) is a computer simulation of the dynamics of a set of particles that are under the influence of a physical force, described by classical physics (Alder and Wainwright 1959; Rahman 1964). Historically, MD simulations have been widely applied to problems involving a potential energy field, such as the gravitational potentials of astrophysics and cosmology (von Hoerner 1963, 1960; Rahman 1964). In the 1970's, MD simulations were first coupled to molecular mechanics (MM) forcefields to investigate protein dynamics.

In MD simulations of biomolecular systems, Newtonian mechanics is used to generate a series of time-dependent conformations of the protein system from a starting conformation (usually an experimentally determined structure) and a set of initial velocities. The potential energy of the system is described as a function of the atoms' position using a molecular mechanics (MM) forcefield. The change in energy and the forces acting on each atom are calculated at each step in the simulation. The time-dependent structural, dynamic and thermodynamic properties of the system can be calculated from the MD trajectories. As discussed earlier, one major limitation of classical MD is the limited sampling of the conformational space of the protein (or protein–ligand) system due to time-dependent trapping in local potential energy minima. Since the 1990s a number of enhanced sampling methods have been developed to address these limitations (Lorenz and Doltsinis 2012; Náray-Szabó et al. 2012; Nowak 2012).

Free energy calculations

One powerful application of MD simulation techniques is the estimation of the change in free energy on the binding of a drug to its target protein. This can, in principle, be directly related to the experimentally determined binding affinity. A series of methods have been developed to effectively sample the intermediate states in the transition between the ligand free in solution and the formation of the protein-ligand complex. The two most commonly used methods are free energy perturbation (FEP) and thermodynamic integration (TI). In FEP, the difference in energy between the dissociated ligand in solution and the ligand-bound complex is used to derive a thermodynamic cycle from which the binding free energy of the ligand can be calculated. In TI, the free energy difference is determined by defining a thermodynamic path between the bound and free state of the ligand. The path chosen is a simplified (mostly one-dimensional) version of the true diffusive path of the ligand binding to the protein.

The two methods have different applications. FEP methods are generally applied to small ligands or drugs with buried binding sites. However, highly charged ligands can be problematic due to their large solvation free energies. TI methods are more applicable for drugs or ligands that have a surface-accessible binding mode or binding site (Gumbart et al. 2012; Jensen 2007; Shirts 2012). Independent of the method used, the task of accurately predicting the absolute binding free energies remains "a daunting computational endeavor" (Gumbart et al. 2012) and are among the most challenging types of biomolecular simulations (Shirts 2012).

Current successes and future directions

The field of rational or computer-aided drug design is built on the premise that knowing the molecular structure and dynamics of a ligand-binding site will enable the design of an optimized ligand molecule that can act as a more effective drug. Structural knowledge of the protein of interest and its dynamics is an essential pre-requisite. For the majority of drug targets, the structure and dynamics of the human protein is unknown. As experimental and modeling techniques improve, highly specific clinical drugs with minimal cross-reactivity (and side effects) are becoming a reality. The carbonic anhydrase inhibitor and anti-glaucoma drug dorzolamide was the first drug approved for clinical use that was developed by computer-aided drug design. This was rapidly followed by HIV protease inhibitors (De Lucca et al. 1997; Greer et al. 1994) and the cancer chemotherapeutic agent and tyrosine kinase inhibitor, imatinib (gleevec) (Druker and Lydon 2000). Current drugs developed by computer-aided drug design target 5-HT3 and acetylcholine receptor agonists, G-protein coupled receptors such as CCR5 and NK1, proton pumps and TRP channels.

Summary

As computational power and modeling algorithms continue to improve, computational modeling is becoming an increasingly important investigative tool. As our experimental knowledge of protein structure increases, so too does the application of computational modeling in elucidating the molecular details of drug binding and the mechanism by which drugs modulate protein structure and function.

Figure legends

Figure 1. An overview of the main techniques used to model proteins and peptides for drug design. The main aims of these techniques is to (i) model the three-dimensional structure of proteins of current or potential drug targets; (ii) identifying and characterizing the structural dynamics associated with the function of a particular protein or peptide; and (iii) predicting the structure and molecular interactions of protein-ligand complexes. In the absence of an experimentally determined protein structure, homology modeling techniques can be used to predict the structure of a given protein from the sequence similarity to homologues of known structure. Docking simulations can be used to predict the interaction of drugs, ligands or other proteins with the protein of interest. From a high-resolution experimentally determined protein structure, the conformational dynamics within the protein and dynamic binding of drug molecules can be elucidated using molecular dynamics simulation techniques.

Figure 2. Molecular docking conformations of protein-protein and protein-ligand docking simulations. Multiple poses of **A)** a pseudokinase domain (colored from orange to red) docked onto a kinase domain (blue); and **B)** an inhibitor molecule (CPK coloring, licorice) docked onto a homology model of a human homologue of LeuT (grey ribbons). Note the extended loops in the model correspond to amino acid insertions not present in the LeuT structure.

REFRENCES

Alder BJ, Wainwright TE (1959) Studies in Molecular Dynamics. I. General Method. Journal of Chemical Physics 31:459-466

Baron R, Nichols SE, McCammon JA (2012) On the use of molecular dynamics receptor conformations for virtual screening. In: Baron R (ed) Computational drug discovery and design. Springer,

Cross AJ, Yocca FD (2012) Essential CNS drug development - pre-clinical development. In: Kalali A, Preskorn S, Kwentus J, Stahl SM (eds) Essential CNS drug development. Cambridge University Press,

De Lucca GV, Erickson-Viitanen S, Lam PYS (1997) Cyclic HIV protease inhibitors capable of displacing the active site structural water molecule. Drug Discovery Today 2:6–18

Druker BJ, Lydon NB (2000) Lessons learned from the development of an Abl tyrosine kinase inhibitor for chronic myelogenous leukemia. Journal of Clinical Investigation 105:3–7

Goodfellow JM (2008) Computer modelling in molecular biology

Greer J, Erickson JW, Baldwin JJ, Varneyl MD (1994) Application of the three-dimensional structures of protein target molecules in structure-based drug design. Medicinal Chemistry 37:1035-1054

Gumbart JC, Roux B, Chipot C (2012) Standard binding free energies from computer simulations: what is the best strategy? Journal of Chemical Theory and Computation 9 (1):794-802. doi:10.1021/ct3008099

Hefti FF (2004) Drug discovery for nervous system diseases. John Wiley & Sons, Jensen F (2007) Introduction to computational chemistry

Kozma D, Simon I, Tusnády G (2013) PDBTM: Protein Data Bank of transmembrane proteins after 8 years. Nucleic Acids Research 41:D524-529

Lorenz C, Doltsinis NL (2012) Molecular dynamics simulation: from "ab initio" to "coarse grained". In: Leszczynski J (ed) Handbook of computational chemistry. Springer,

Náray-Szabó G, Perczel A, Láng A (2012) Protein modeling. In: Leszczynski J (ed) Handbook of computational chemistry. Springer,

The Nobel Prize in Chemistry 2013 - Press Release. (2013) Nobel Media AB. http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2013/press.html.

Nowak W (2012) Applications of computational methods to simulations of proteins dynamics. In: Leszczynski J (ed) Handbook of computational chemistry. Springer,

Rahman A (1964) Correlations in the Motion of Atoms in Liquid Argon. Physical Review 136:405-411

Rangwala H (2010) A survey of remote homology detection and fold recognition methods. In: Rangwala H, Karypis G (eds) Introduction to protein structure prediction methods and algorithms. Wiley,

Rost B (1999) Twilight zone of protein sequence alignments. Protein Engineering, design & selection 12:85-94

RSCB Protein Data Bank. Research Collaboratory for Structural Bioinformatics. http://www.rcsb.org/pdb/home/home.do. Accessed 10/12/2013

Schushan M, Ben-Tal N (2010) Modeling and validation of transmembrane protein structures. In: Rangwala H, Karypis G (eds) Introduction to protein structure prediction methods and algorithms. Wiley,

Shirts MR (2012) Best practices in free energy calculations for drug design. In: Baron R (ed) Computational drug discovery and design. Springer,

Squire LR, Berg D, Bloom FE, du Lac S, Ghosh A, Spitzer NC (2008) Fundamental neuroscience. Burlington Elsevier,

von Hoerner S (1960) Die numerische Integration des n-Körper-Problemes für Sternhaufen, I. Zeitschrift für Astrophysik 50:184-214

von Hoerner S (1963) Die numerische Integration des n-Körper-Problems für Sternhaufen, II. Zeitschrift für Astrophysik 57:47-82

Xu Y, Xu D, Liang J, Wooley JC (2007) Computational methods for protein structure prediction and modeling. Springer,

Zvelebil M, Baum JO (2008) Understanding bioinformatics. Garland Science,