1. Discuss the molecular dynamics method.

By analysing a molecular system using physics, specifically the laws of motion and energy, one can model the movements of atoms within a 3d space or medium. By treating all atoms as 'balls' the motions can be calculated based on their interactions with other atoms. These interactions are derived using four main atomic interactions which can be expressed as a 'spring model'. So we can simulate each state from beginning to the end of the interaction. All interactions are assigned a spring constant (k) which is the energy required to push them past equilibrium. The spring constant is denoted as the equilibrium as two atoms join at rest.

- **Bond stretching**

  The energy required to pull them apart or push them together can be expressed with the equation.

  $$\sum k_b \, (r - r_o)^2$$

  k_b is the spring constant
  r_o is the bond length at equilibrium

- **Angle bonding**

  Energy required to bend a bond

  $$\sum k_\theta \, (\theta - \theta_o)^2$$

  k_θ is the spring constant on the bend
  θ_o is the bond length at equilibrium

-
- **Torsion**

  The energy required to twist a bond

  $$\sum A \, [\, 1 + \cos(n\tau - \phi)]$$

  A controls amplitude
  n controls periodicity, ϕ shifts the entire curve along the axis τ

- **Non covalent bonds**

  Describes how atoms behave when no bond is present using electrostatic and Van Der Waal equations

  $$E = \sum_i \sum_j \frac{-A_{ij}}{r_{ij}^6} + \frac{B_{ij}}{r_{ij}^{12}} + \sum_i \sum_j \frac{q_i q_j}{r_{ij}}$$

  A determines the degree of attractiveness
  B determines the degree of repulsion
  Q is the charge

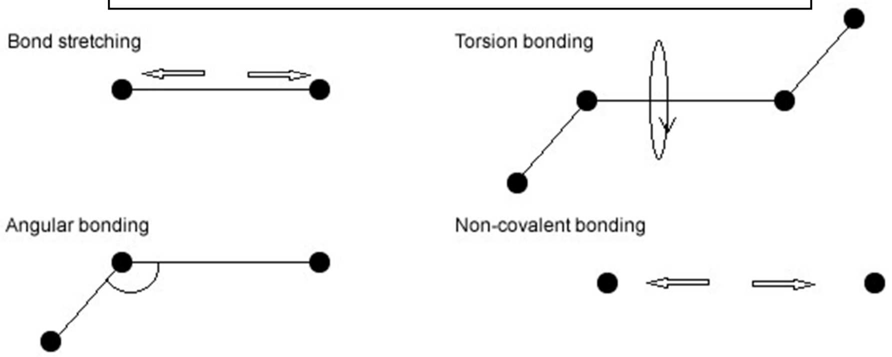


Figure 1 Four bond interactions of a molecules

This is an ab-initio (from the start) method, which uses only calculations and no previous information, as opposed to using existing databases and knowledge.

By mapping these movements using chemical bond attraction values, using covalent, ionic or electrostatic, we can simulate the movements of atoms as a group and hence predict the movements of a compound. These values are generally known as a 'forcefield' and are developed using both experimental and empirical quantum mechanics. In the case of protein structure predication the amino acid compounds are modelled in this way and the aim is to predict where they will end up based on a primary sequence derived from DNA.

Using the energetic interactions between atoms sequential states of protein development can be found over a given time period.

Rugged energy landscape

The amount of states a sequence can take while forming from 150 amino acids is approximately 10^300 but proteins do it, on average, parsing only 10^8 states, so some other force is a play. Something drives the amino acid chain to achieve its native state i.e. a desired protein. This can be modelled as a 'rugged energy landscape' based on the observation that the states seem to close in on the native state. The "way" or "path" can be defined as the states the protein takes while moving towards its native state. In reality there are many paths a protein can take but all are guided by achieving a lower energy state. So if we envisage this energy landscape as a mountain range the protein creation process always tries to get to lower ground. Obviously if there is a valley the process will need more energy to get out of it and continue on its way. The entire process (around 10^8 states) is achieved in around 2-3 seconds in biology, but would take many years using modern computational methods.

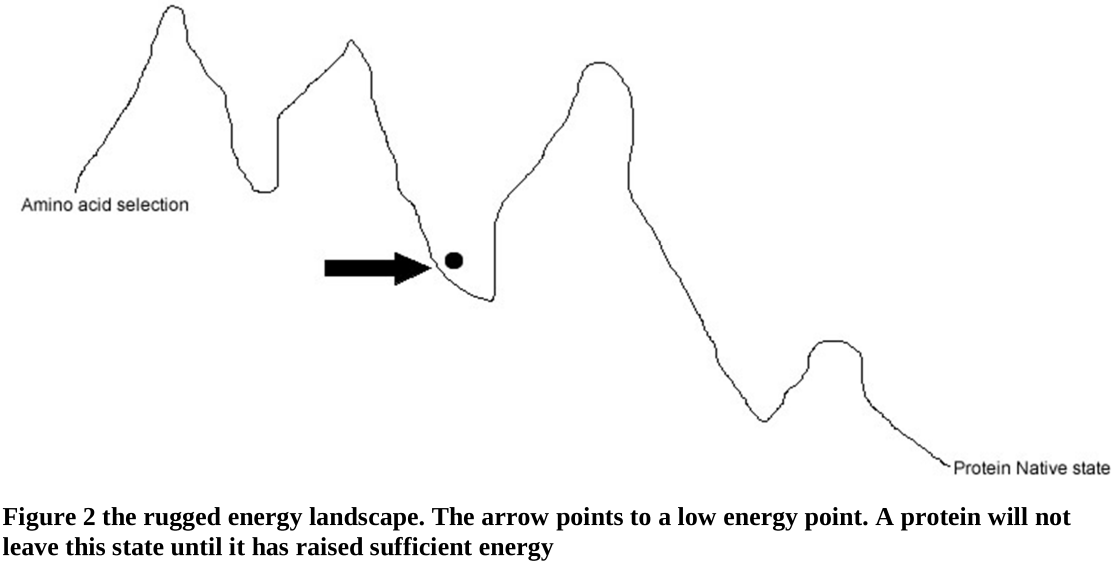


Figure 2 the rugged energy landscape. The arrow points to a low energy point. A protein will not leave this state until it has raised sufficient energy

The problem of simulation time exists due to restrictions in computing power; various methods have been developed to speed up a simulation.

- **Energy Flooding**

  When using computer methods we can speed them up by adding energy to the process whenever a state is "stuck" in a valley. We must ensure we monitor the amount of energy we add and when we add it, so we can reconstruct the energy landscape.
- **Removal of solvent**

  In reality all reactions will take place in a solvent, by incorporating this into the simulation this will increase the amount of interactions between atoms, and hence increase time and calculations necessary. Removing this will increase the speed but decrease the accuracy of the simulation.

- **Periodic boundary conditions**

  The largest cost of the solvent in a simulation is the edge water i.e. calculating interactions of the water molecules on the edge. By modelling the reaction in a cube we can create an infinite space. If a water molecule gets to the edge it appears on the other side, thus removing edge water and minimising the amount of solvent reactions.

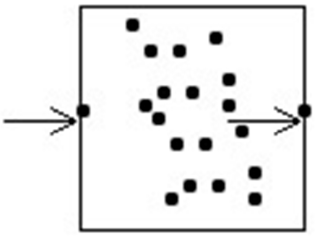


Figure 3 an illustration of periodic boundary conditions

- **Granular methods**

  Instead of calculating the movements of each atom we can group atoms into larger structures thus reducing the amount of calculations. So a group of 9 could be shown and treated as just one or two molecules.

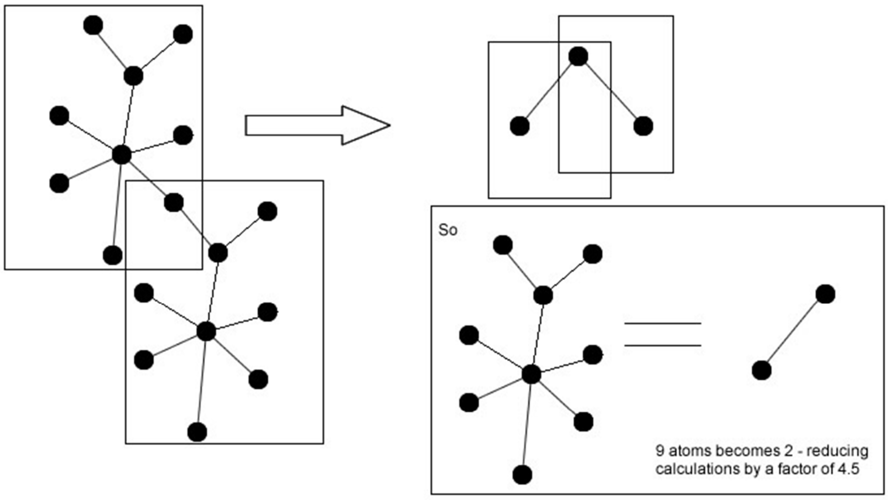

So
9 atoms becomes 2 - reducing calculations by a factor of 4.5

Figure 4 Course grained methods

- **Steered molecular dynamics**
  A force is applied to an atom or system and the force required to "steer it" or achieve desired results is noted. E.g. when simulating membrane permeation an atom is held in the membrane and the requiring force is noted. The higher the force required the less it wants to be there and hence the less it is likely to go there on its own.

Even using these methods and with improving computing powers we cannot find the tertiary structure of an average size protein we can only simulate a fraction of the process, around 1-2 Nano-seconds.

Molecular dynamics can be used in drug development by simulating ligand docking and cellular membrane permeation

Because cell membranes, proteins and drugs are all atomic compounds we can use the above methods to simulate the interactions between these compounds. We can model how a drug will dock to a protein or enzyme as well as whether a drug will penetrate a cell membrane.

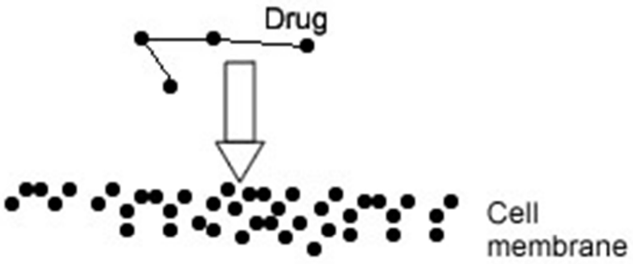


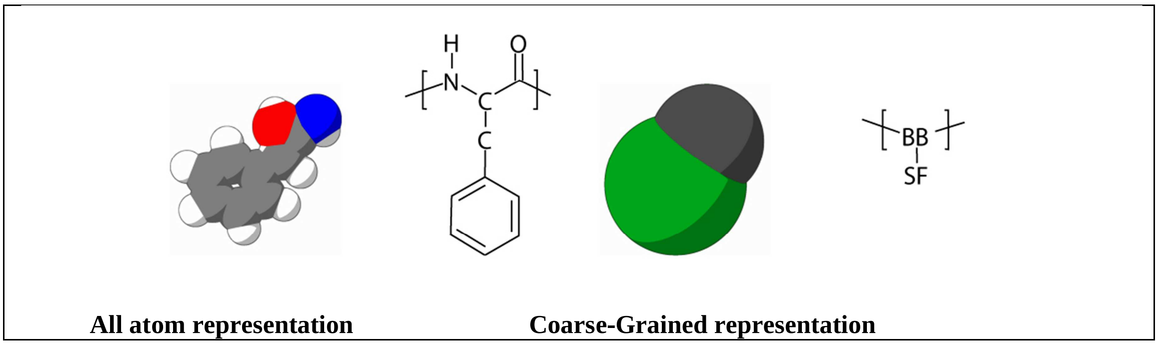
Figure 5 A drug molecule permeating through a membrane. This illustration is at an atomic level.

2. Perform a literature search, identify and discuss 2 recent papers (later than 2010) describing the use of MD simulations to study biological problems.

***Journal article:* Coarse Grained Molecular Dynamics Simulations of Transmembrane Protein-Lipid Systems**

The method discussed in this journal article is aimed at reducing the amount of calculations required to perform a simulation. This journal article is mainly interested in the setup and analysis of this simulation technique but does outline some uses for it.

To implement a Coarse Grained method they map the original identified atoms involved in the simulation to Coarse grained (CG) particles. In this approach they have grouped four non-hydrogen atoms into one CG particle. To keep the interactions correct the CG modules are mapped to the amino acids in two different ways, one for the side chain and one for the backbone. For atomic interactions the Van der Waals radius for the CG group is calculated from its volume.



Figures 8 Coarse Grained translation (Spijker *et al 2010, p 4)*

All covalent values, bond stretching and angular are calculated from MD simulations of several nano-seconds and a probability distribution model is calculated. Torsional interactions have been omitted.

Three non-bonded interactions are stipulated

- Protein-Protein
- Environment-Protein
- Environment-Environment

For interactions between water and lipid (involving Environment) the Markvoort's forcefield values are used. All other interactions are used based on various MD principles and data.

Two different sets of MDs are performed environment-protein and protein-protein. For the former each protein particle is pulled through a lipid membrane and the 'steering technique' (force needed to pull protein through is calculated) is used. For the protein-protein MDs five large proteins are simulated in 5 separate simulations for 60ns. These five are mapped and then translated into their Coarse Grained equivalents before simulation. This is to gauge the behaviour of the CG modules in larger systems.

In this work the coarse grained interaction behaviour is based largely on existing data with some interaction calculations introduced. Unlike many other CG models the torsional interaction potentials were ignored. Using the tests above they were able to show that the main (most influential) parts of this model behaved as expected. As mentioned the Journal article outlines some applicable uses for this method namely for study of WALP-peptides to simulate interactions between proteins and lipids

Journal: Unfolding of the Amyloid β-Peptide Central Helix: Mechanistic Insights from

Molecular Dynamics Simulations

One of the identified causes of the Alzheimer's disease (AD) is the polymerization of the Amyloid Beta-peptide (Aβ) (fig 6), which requires unfolding (fig 7) from its native state. This causes toxic assemblies of (Aβ). In this work molecular dynamic simulations were used to examine how these mutant states of Aβ were achieved.

This unfolding was done using the following three-step mechanism:

- loss of helical backbone hydrogen bonds
- strong interactions between nonpolar sidechains
- strong interactions between polar sidechains.

If any of theses steps are omitted then the protein does not fully unfold, hence a drug that could replicate this could be developed to inhibit unfolding and hence prevent Alzheimer's.

Molecular Dynamics

The protein Aβ contains between 36 and 43 amino acids, the varying amount is due to different isoforms. The most common isoform is Aβ40 but Aβ42 is more commonly associated with Alzheimer's. It was found that the middle part of Aβ42 (15-24) forms an α-helix conformation in membrane-water type mediums, this has a sequence of HHQKLVFFAEDGS. It is the unfolding of this that forms the amyloid fibril that causes AD. The protein that was used in these simulations was Wild Type Aβ 13-26 and the two mutant forms alanine (MA) and luecine (ML) were substituted at the 3 non-polar sites.

- The simulations were carried out using the CHARM22/CMAP fourcefield, they also used the SHAKE algorithm, and visualisations of the simulations were modelled using VMD software.
- The edge solvent loss was solved by using periodic boundary conditions.
- The temperature was checked every 4 ps and was found to be within 5K of the target bath temperature.
- Time was spent improving the lookup process for non-covalent bonding values to speed up the simulations
- No harmonic restraints were imposed on any molecules

The coordinates were saved every 1ps and the entire simulation went for 20ns.

Due to the time that this simulation was run for (20ns) there was an optimal temperature that the unfolding of Aβ would take place (within that 20ns). To achieve this, simulations were done at 300K, 330K and 360K and the results were recorded.

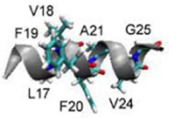


Fig 6 Amytoid Beta-peptide

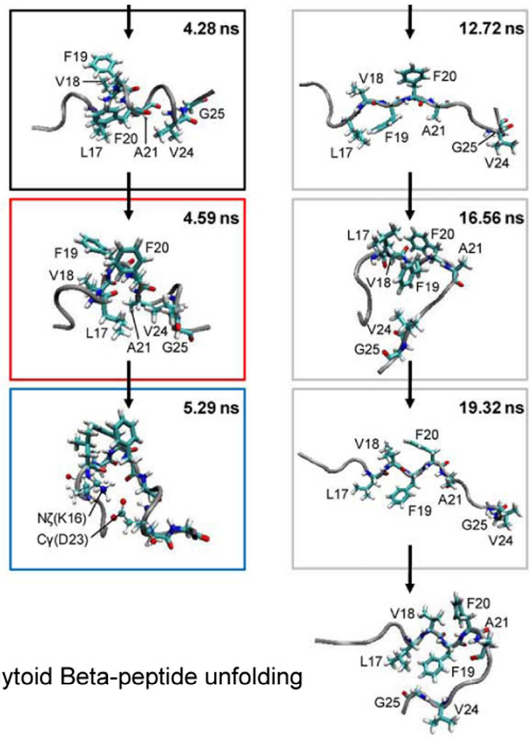


Fig 7 Amytoid Beta-peptide unfolding

Figures 6, 7 *(Ito et al 2011, p 5)*

In this study existing knowledge of the cause of AD was used, that is that the unfolding of the Amyloid Beta peptide forms the amyliod fibril that causes Alzheimer's. They then simulated this unfolding using molecular dynamics and monitored and recorded this, this on its own was still not useful. They then identified steps in the unfolding process that are required to complete the unfolding and omitted these and re-ran the simulation. By doing this they found events and conditions that could be replicated to prevent unfolding of Aβ. Now, potentially, drugs can be created that re-produce these events, thus preventing the toxic formations that cause AD.

3. *Perform a literature search, and discuss in detail one computational method developed to study protein folding (examples include conformational flooding and coarse-grained methods). In particular, discuss its advantages, disadvantages, and how you think it can be improved*

Journal article: Active learning for human protein-protein interaction prediction

Summary

Determining whether a pair of proteins interact, experimentally, is resource-intensive. Training a machine can guide the selection of interacting protein pairs. The concept discussed in this Journal article is the training of such systems to reduce the required number of interactions that need to be calculated. In summary a program is run on a collection of current known interacting and non-interacting protein pairs and a database is populated thus eliminating these interactions from future calculations. This collection is known as the human protein interactome.

Data and comparison methodology

The dataset used in this example contained 14600 known interaction protein pairs (positive pairs) and 400,000 randomly selected non-interacting pairs (random pairs). Each pair is assigned a feature vector containing the following information, from Gene Oncology db, BLASTP and NCBI (and others):

- Gene Oncology (GO) cell component
- Gene Oncology molecular function
- Gene Oncology biological process
- Co-occurrence in tissue
- Gene expression
- Homology based sequence similarity
- Domain interaction

A random forest (rf) of decision trees is then trained by iterating through and assigning labels to each node. This is done by iterating and then adding to the labels based on information gleaned from their information vectors. So a positve pair's (known interacting pair) label contains more information then a random pair.

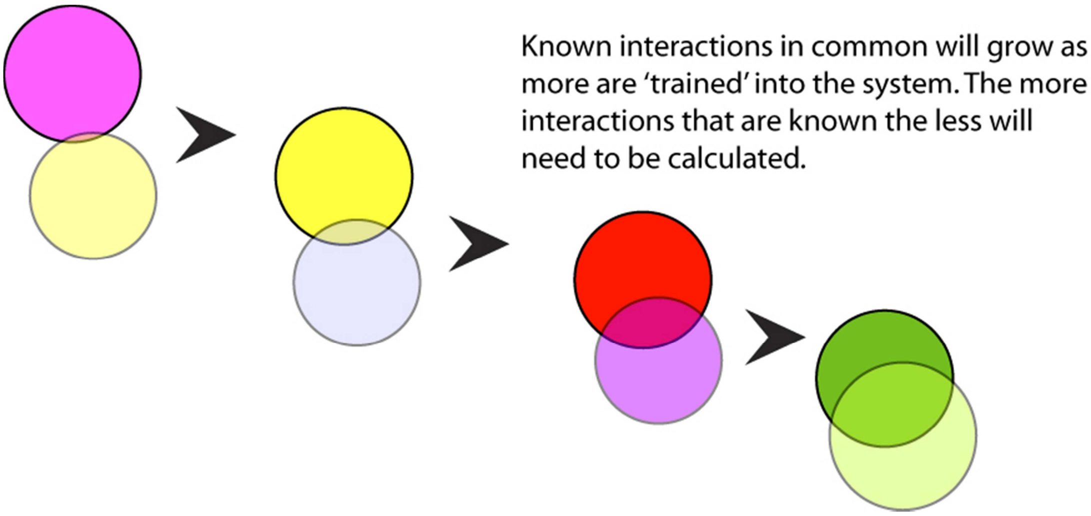
Discussion

Quite simply the idea here and in other uses of this technology is to ‘train the system’. In artificial intelligence (AI) this notion is usually based around teaching the system to make decisions based on some data set. Whilst this is partially true in this case the main purpose is to reduce the number of calculations, so in effect the goal is to use previous calculations rather then repeat them. In essence, if an interaction has been calculated in the past why do it again? The challenge being to teach a system to recognise these past calculations, to do this a simulation must be able to identify an interaction as previously occuring, hence the training. In AI one of the main methods used is inference. E.g:

1. All humans have hair
2. Harry has hair
3. Therefore Harry is a human

Obviously other things have hair that are not humans so while this is accurate, there is the case of sub-setting, which is - not including all options in a predicate. In this work the predicates are protein interactions, or pairs, so we need to realise that all interacting training pairs used must be accurate to assume the correct outcome.

This could be re-purposed for protein structure prediction by recording all interactional data so that particular calculation is not required next time. It would be less effective for this use as the sheer number of molecular interactions involved in forming a protein is much higher then interactions between, say, 200 proteins. In the example below is a group of systems with some calculations in common. As you can see a system can be trainined to exponentially improve its self.



Known interactions in common will grow as more are 'trained' into the system. The more interactions that are known the less will need to be calculated.

Figure 9 A system improving itself, the intersects are known interactions between to systems

Reference

Mika Ito, Jan Johansson, Roger Strömberg, Lennart Nilsson , 2011, ***Unfolding of the Amyloid β-Peptide Central Helix: Mechanistic Insights from Molecular Dynamics Simulations***, Journal Article, ***PLoS ONE*** 6 (3) p. 13, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3049775/?tool=pmcentrez&rendertype=abstract

Peter Spijker, Bram Van Hoof, Michel Debertrand, Albert J Markvoort, Nagarajan Vaidehi, Peter A J Hilbers, 2010, ***Coarse Grained Molecular Dynamics Simulations of Transmembrane Protein-Lipid Systems***, Journal Article, ***International Journal of Molecular Sciences*** 11 (6) p. 2393-2420 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2904924/?tool=pmcentrez&rendertype=abstract

Thahir P Mohamed, Jaime G Carbonell, Madhavi K Ganapathiraju, 2010, ***Active learning for human protein-protein interaction prediction***, Journal Article, ***BMC Bioinformatics*** 11 (Suppl 1) p. S57, http://www.ncbi.nlm.nih.gov/pubmed/20122232

Q1) 3
Q2) 8
Q3) 7

Total: 18/20