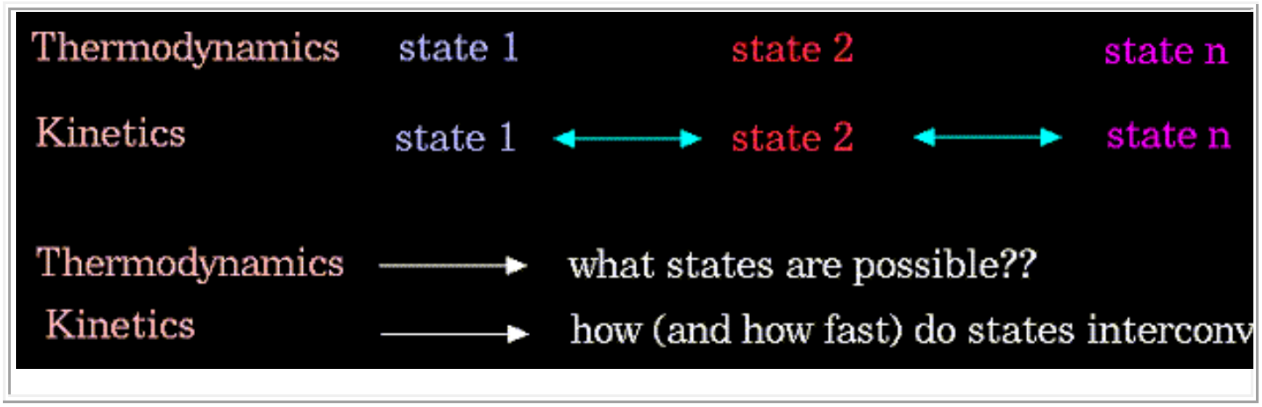
Molecular Dynamics

In the broadest sense, molecular dynamics is concerned with molecular motion. Motion is inherent to all chemical processes. Simple vibrations, like bond stretching and angle bending, give rise to IR spectra.

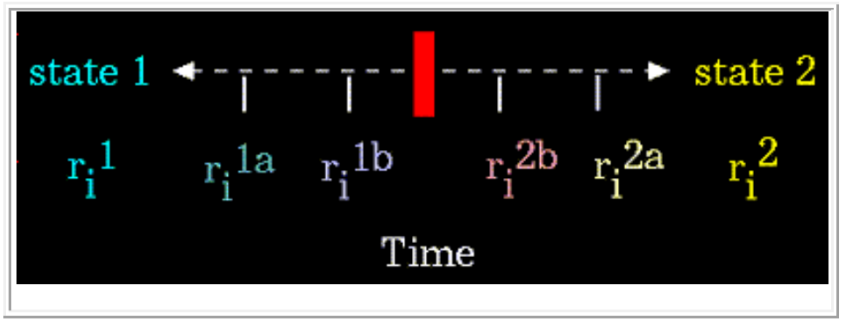
Chemical reactions, hormone-receptor binding, and other complex processes are associated with many kinds of intra- and intermolecular motions.

The **driving force** for chemical processes is described by **thermodynamics**. The **mechanism** by which chemical processes occur is described by **kinetics**. Thermodynamics dictates the energetic relationships between different chemical states, whereas the sequence or rate of events that occur as molecules transform between their various possible states is described by kinetics:



Conformational transitions and local vibrations are the usual subjects of molecular dynamics studies. Molecular dynamics alters the intramolecular degrees of freedom in a step-wise fashion, analogous to energy minimization. The individual steps in energy minimization are merely directed at establishing a down-hill direction to a minimum. The steps in molecular dynamics, on the other hand, meaningfully represent the changes in atomic position, r*i*, over time (i.e. velocity).

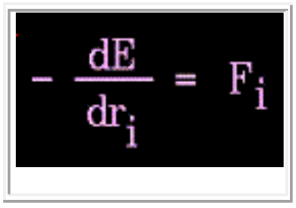
For the "i" atoms of the system:



Newton's equation is used in the molecular dynamics formalism to simulate atomic motion:

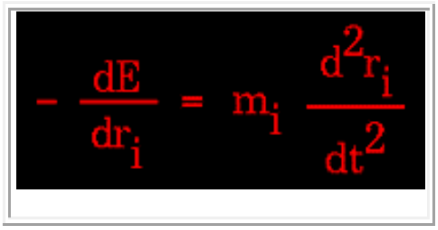
The rate and direction of motion (velocity) are governed by the forces that the atoms of the system exert on each other as described by Newton's equation. In practice, the atoms are assigned initial velocities that conform to the total kinetic energy of the system, which in turn, is dictated by the desired simulation temperature. This is carried out by slowly "heating" the system (initially at absolute zero) and then allowing the energy to equilibrate among the constituent atoms. The basic ingredients of molecular dynamics are the calculation of the force on each atom, and from that information, the position of each atom throughout a specified period of time (typically on the order of picoseconds = 10^-12 seconds).

The force on an atom can be calculated from the change in energy between its current position and its position a small distance away. This can be recognized as the derivative of the energy with respect to the change in the atom's position:

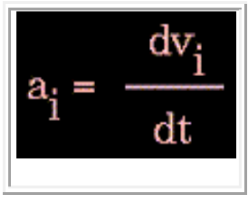


Energies can be calculated using either molecular mechanics or quantum mechanics methods. Molecular mechanics energies are limited to applications that do not involve drastic changes in electronic structure such as bond making/breaking. Quantum mechanical energies can be used to study dynamic processes involving chemical changes. The latter technique is extremely novel, and of limited availability (Gaussian03 is an example of such a program).

Knowledge of the atomic forces and masses can then be used to solve for the positions of each atom along a series of extremely small time steps (on the order of femtoseconds = 10^-15 seconds). The resulting series of snapshots of structural changes over time is called a trajectory. The use of this method to compute trajectories can be more easily seen when Newton's equation is expressed in the following form:



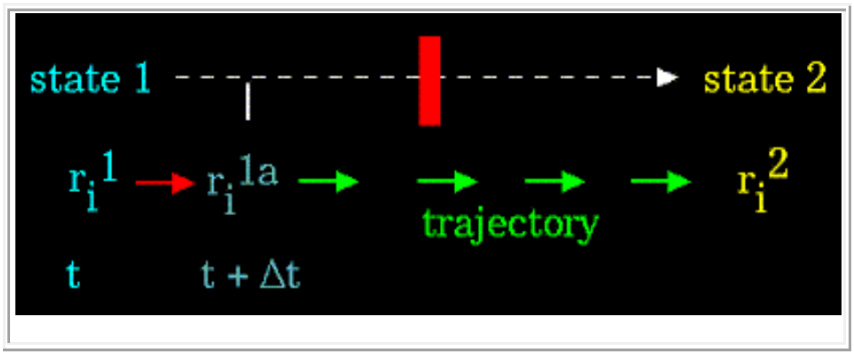
In practice, trajectories are not directly obtained from Newton's equation due to lack of an analytical solution. First, the atomic accelerations are computed from the forces and masses. The velocities are next calculated from the accelerations based on the following relationship:



Lastly, the positions are calculated from the velocities:

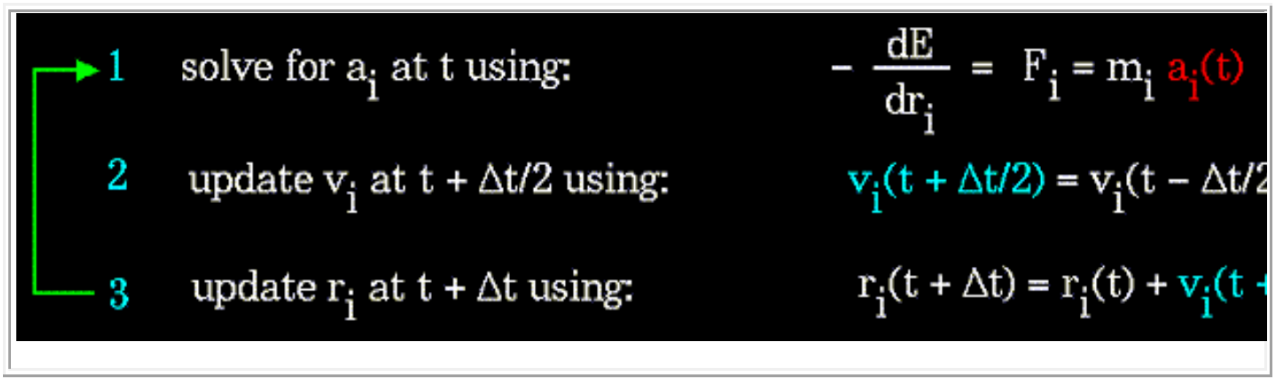
$$v_i = \frac{dr_i}{dt}$$

A trajectory between two states can be subdivided into a series of sub-states separated by a small time step, "delta t" (e.g. 1 femtosecond):



The initial atomic positions at time "t" are used to predict the atomic positions at time "t + delta t". The positions at "t + delta t" are used to predict the positions at "t + 2*delta t", and so on.

The "leapfrog" method is a common numerical approach to calculating trajectories based on Newton's equation. The steps can be summarized as follows:



The method derives its name from the fact that the velocity and position information successively alternate at 1/2 time step intervals.

Molecular dynamics has no defined point of termination other than the amount of time that can be practically covered. Unfortunately, the current picosecond order of magnitude limit is often not long enough to follow many kinds of state to state transformations, such as large conformational transitions in proteins.

Molecular dynamics calculations can be performed using both HyperChem and Gaussian programs.

Quantum mechanics:

Definition of Computational Chemistry

- Computational Chemistry: Use mathematical approximations and computer programs to obtain results relative to chemical problems.
- Computational *Quantum* Chemistry: Focuses specifically on equations and approximations derived from the postulates of quantum mechanics. Solve the Schrödinger equation for molecular systems.
- *Ab Initio* Quantum Chemistry: Uses methods that do not include any empirical parameters or experimental data.

What's it Good For?

- Computational chemistry is a rapidly growing field in chemistry.
  - Computers are getting faster.
  - Algorithims and programs are maturing.
- Some of the almost limitless properties that can be calculated with computational chemistry are:
  - Equilibrium and transition-state structures
  - dipole and quadrapole moments and polarizabilities
  - Vibrational frequencies, IR and Raman Spectra
  - NMR spectra
  - Electronic excitations and UV spectra
  - Reaction rates and cross sections
  - thermochemical data

Motivation

- Schrödinger Equation can only be solved exactly for simple systems.
  - Rigid Rotor, Harmonic Oscillator, Particle in a Box, Hydrogen Atom

- For more complex systems (i.e. many electron atoms/molecules) we need to make some simplifying assumptions/approximations and solve it numerically.
- However, it is still possible to get very accurate results (and also get very crummy results).
  - In general, the "cost" of the calculation increases with the accuracy of the calculation and the size of the system.

Getting into the theory...

- Three parts to solving the Schrödinger equation for molecules:
  - Born-Oppenheimer Approximation
    - Leads to the idea of a potential energy surface
  - The expansion of the many-electron wave function in terms of Slater determinants.
    - Often called the "Method"
  - Representation of Slater determinants by molecular orbitals, which are linear combinations of atomic-like-orbital functions.
    - The basis set

The Born-Oppenheimer Approximation

- Now we can solve the electronic part of the Schrödinger equation separately.

$$\hat{H}_{el}\psi_{el}(r;R) = E_{el}\psi_{el}(r;R)$$

$$\hat{H}_{el} = -\frac{\hbar^2}{2m_e}\sum_i \nabla_i^2 - \sum_\alpha \sum_i \frac{Z_\alpha e'^2}{r_{i\alpha}} + \sum_j \sum_{i>j} \frac{e'^2}{r_{ij}}$$

- BO approximation leads to the idea of a potential energy surface.

$$U(R) = E_{el} + V_{NN}$$

$$V_{NN} = \sum_\alpha \sum_{\alpha>\beta} \frac{Z_\alpha Z_\beta e'^2}{r_{\alpha\beta}}$$

Nuclear Schrödinger Equation

Once we have the Potential Energy Surface (PES) we can solve the nuclear Schrödinger equation.

- Solution of the nuclear SE allow us to determine a large variety of molecular properties.

  An example are vibrational energy levels.

Energy Minimization

The potential energy calculated by summing the energies of various interactions is a numerical value for a single conformation. This number can be used to evaluate a particular conformation, but it may not be a useful measure of a conformation because it can be dominated by a few bad interactions. For instance, a large molecule with an excellent conformation for nearly all atoms can have a large overall energy because of a single bad interaction, for instance two atoms too near each other in space and having a huge van der Waals repulsion energy. It is often preferable to carry out energy minimization on a conformation to find the best nearby conformation. Energy minimization is usually performed by gradient optimization: atoms are moved so as to reduce the net forces on them. The minimized structure has small forces on each atom and therefore serves as an excellent starting point for molecular dynamics simulations.

It is also possible to minimize the energy of a conformation by optimizing the dihedral angle degrees of freedom, rather than the Cartesian coordinates. The minimization occurs in -dimensional space, where is the number of dihedral angles. Torques, or derivatives of the forcefield with respect to dihedral angles, take the place of the gradient. We have found that ``torque minimization," when followed by Cartesian minimization, produces an overall lower-energy conformation than Cartesian minimization alone. Neither method, however, can guarantee that the lowest possible conformation (the global minimum) will be reached. The process of moving along pathways in conformational space usually ends at a ``local minimum" - a well in the potential energy surface, where the energy is lower than for all other nearby conformations, but not necessarily lower than other local minima.

De Novo Protein Design

In rational protein design proteins can be redesigned from the sequence and structure of a known protein, or completely from scratch in *de novo* protein design. In protein redesign, most of the residues in the sequence are maintained as their wild-type amino-acid while a few are allowed to mutate. In *de novo* design the entire sequence is designed anew, based on no previous sequence.

Both *de novo* designs and protein redesigns can establish rules on the sequence space: the specific amino acids that are allowed at each mutable residue position. For example, the composition of the surface of the RSC3 probe to select HIV-broadly neutralizing antibodies was restricted based on evolutionary data and charge balancing. In fact, many of the earliest attempts on protein design were heavily based on empirical "rules" on the sequence space. Moreover, the design of fibrous proteins, usually follows strict rules on the sequence space. Collagen-based designed proteins, for example, are often composed of Gly-Pro-X repeating patterns. With the advent of computational techniques, however, the design of proteins with no human intervention in sequence selection has become possible.

PROTEIN DATABASES

Protein databases are more specialized than primary sequence databases.

They contain information derived from the primary sequence databases.

Some contain protein translations of the nucleic acid sequences.

Some contain sets of patterns and motifs derived from sequence homologs.

GenBank - the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences.

PIR Protein Information Resource -a comprehensive, non-redundant, expertly annotated, fully classified and extensively cross-referenced protein sequence database.

SWISS-PROT &TrEMBL - SWISS-PROT is a curated protein sequence database. is a computer-annotated supplement of SWISS-PROT that contains all the translations of EMBL nucleotide sequence entries not yet integrated in SWISS-PROT.

TIGR - a collection of curated databases containing DNA and protein sequence, gene expression, cellular **role, protein family, and taxonomic data for microbes, plants and humans.**

MOTIF, PATTERN & PROFILE DATABASES

ALIGN - a compendium of sequence alignments: it is a companion resource to *PRINTS*.

BLOCKS - multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins.

DOMO - a database of homologous protein domain families.

HOMSTRAD - a curated database of structure-based alignments for homologous protein families.

InterPro- Integrated Resource of Protein Domains and Functional Sites - InterPro is an integrated documentation resource for protein families, domains and sites, developed initially as a means of rationalising the complementary efforts of the PROSITE, PRINTS, Pfam and ProDom database

projects. Each combined InterPro entry includes functional descriptions and literature references, and links are made back to the relevant member database(s), allowing users to see at a glance whether a particular family or domain has associated patterns, profiles, fingerprints, etc. Merged and individual entries (i.e., those that have no counterpart in the companion resources) are assigned unique accession numbers. Each InterPro entry lists all the matches against SWISS-PROT and TrEMBL (more than 1,000,000 hits in total).InterPro aims to reduce duplication of effort in the labour-intensive, rate-limiting process of annotation, and will facilitate communication between the disparate resources. By uniting these databases, we capitalise on their individual strengths, producing a single entity that is far greater than the sum of its parts.

PFam - a database of multiple alignments of protein domains or conserved protein regions. The alignments represent some evolutionary conserved structure which has implications for the protein's function. Profile hidden Markov models (profile HMMs) built from the Pfam alignments can be very useful for automatically recognizing that a new protein belongs to an existing protein family, even if the homology is weak.

PRINTS ñ Protein Fingerprint Database - a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family.

PRINTS-S ñ relational cousin of the PRINTS Database

ProDom - an automatic compilation of homologous domains.ProDom families were generated automatically using PSI-BLAST with a profile built from the seed alignments of Pfam-A 4.3 families.

ProSite - is a database of protein families and domains

consisting of biologically significant sites, patterns and profiles.

Protein Profiles - online cross-references to the Oxford University Press Protein Profiles project.

ProtoMap - site offers an exhaustive classification of all the proteins in the SWISSPROT and TrEMBL databases, into groups of related proteins.The resulting classification splits the protein space into well defined groups of proteins, most of them are closely correlated with natural biological families and superfamiliesfor comprehensive evaluation results). The hierarchical organization may help to detect finer subfamilies that make up known families of proteins as well as interesting relations between protein families.

SBASE - protein domain library sequences that contains 237.937 annotated structural, functional, ligand-binding and topogenic segments of proteins, cross-referenced to all major sequence databases and sequence pattern collections.

SYSTERS - SYSTERS cluster set contains sequences from SWISS-PROT ,TrEMBL, PIR, Wormpep, and MIPS Yeast protein translations which are sorted into disjoint clusters. fragmental sequences build single sequence clusters, while the remaining sequences are contained in clusters of non-redundant sequences per cluster.

PROTEIN STRUCTURE DATABASES

CATH Protein Structure Classification ñ a hierarchical domain classification of protein structures in the Brookhaven protein databank.

FSSP Fold Classification based on Structure-Structure Alignment of Proteins - based on exhaustive all-against-all 3D structure comparison of protein structures currently in the Protein Data Bank (PDB).

Library of Protein Family Cores - structural alignments of protein families and computed average core structures for each family.Useful for building models, threading, and exploratory analysis.

ModBase a database of three-dimensional protein models calculated bycomparative modeling.

PRESAGE - a database of proteins, each of which has a collection of annotations reflecting current experimental status, structural assignments models, and suggestions.

RCSB Protein Data Bank - single international repository for the processing and distribution of 3-D macromolecular structure data primarily determined experimentally.

Protein Loop Classification - Conformational clusters and consensus sequences for protein loops derived by computational analysis of their structures.

SCOP ñ Structural Classification of Proteins - a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known.

Sloop Database ñ Sloop Database of Super Secondary Fragments - a classification of protein loops.

3 Dee ñ Database of Protein Domain Definitions - contains structural domain definitions for all protein chains in the Protein Databank (PDB)that have 20 or more residues and are not theoretical models.

References

1. Molecular Dynamics With Deterministic and Stochastic Numerical Methods by Ben Leimkuhler and Charles Matthews

2. Replica-exchange molecular dynamics method for protein folding by Yuji Sugita and Yuko Okamoto

3. The Art of Molecular Dynamics Simulation 2nd Edition by D. C. Rapaport

4. Molecular Dynamics, Volume 7 (1999) 1st Edition From Classical to Quantum Methods - Perla Balbuena Jorge Seminario. eBook ISBN: 9780080536842