Incorporation of Noncanonical Amino Acids into Engineered Proteins

There are two generic (and complementary) strategies for metabolic incorporation of noncanonical amino acids into proteins – the so-called residue-specific and site-specific methods. The residue-specific approach involves replacement of all (or a fraction) of one of the natural amino acid residues. This method has its origins in the work of Cohen and coworkers, who showed in the 1950s that near-quantitative replacement of methionine by selenomethionine could be accomplished in bacterial cells. This observation has had revolutionary consequences for protein science and engineering, in that it provides the basis of the multiwavelength anomalous diffraction method for crystallographic structure determination.

The site-specific approach allows replacement of a single amino acid residue by a noncanonical analog. In this approach, a heterologous transfer RNA(tRNA)/aminoacyl-tRNAsynthetase pair is used to

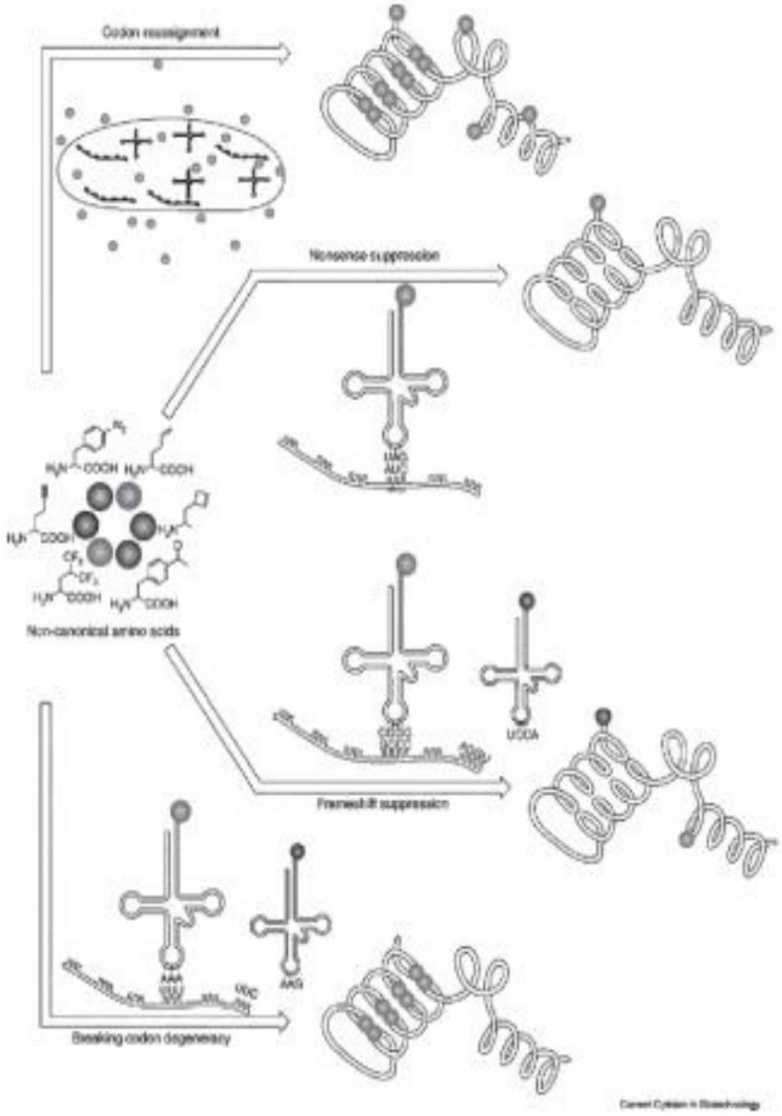


Fig. 1 Methods for incorporation of noncanonical amino acids. Residue-specific incorporation by sense codon reassignment enables replacement of all, or a fraction, of the corresponding canonical residues. Nonsense suppression, frameshift suppression, and breaking codon degeneracy can all be used to place noncanonical amino acids at specific sites. (Reprinted from Link et al. 2003; with permission from Elsevier)

deliver the analog in response to a nonsense or four-base codon. In 1996, Drabkin and coworkers used an Escherichia coli tRNA/glutaminyl-tRNAsynthetase pair for amber codon suppression in mammalian cells, and showed that the suppressor tRNA was not charged by any of the mammalian aminoacyl-tRNAsynthetases. Shortly thereafter, Furter (1998) introduced a yeast tRNA/phenylalanyl-tRNAsynthetase (PheRS) Noncanonical Amino Acids in Protein Science and Engineering 129 pair into E. coli for site-specific incorporation of the noncanonical amino acid p-fluorophenylalanine. Since then, amber codon suppression has become the most common method for site-specific incorporation of noncanonical amino acids in vivo. Schultz and coworkers have been especially successful in producing orthogonal suppressor tRNA/aminoacyl-tRNAsynthetase pairs for incorporation of chemically, structurally, and spectroscopically diverse amino acid analogs. Site-specific incorporation has also been accomplished in Xenopus oocytes using microinjected messenger RNAs and chemically misacylated amber suppressor tRNAs.

Sisido have pioneered the use of four-base codons (frameshift suppression) for site-specific introduction of noncanonical amino acids into proteins, and have employed this strategy to label streptavidin with fluorophores for fluorescence resonance energy transfer (FRET) experiments. Much of the work reported to date with four-base codons involves in vitro translation, but design of appropriate orthogonal tRNA/aminoacyl-tRNAsynthetase pairs enables use of the method in bacterial cells. Anderson and coworkers have reported orthogonal tRNA/leucyl-tRNAsynthetase (LeuRS) pairs for four-base, amber, and opal suppression. Anderson have reported use of a four-base codon with an amber codon for incorporation of two noncanonical amino acids into a recombinant protein using two orthogonal sets. An analogous five-base codon strategy has also been described.

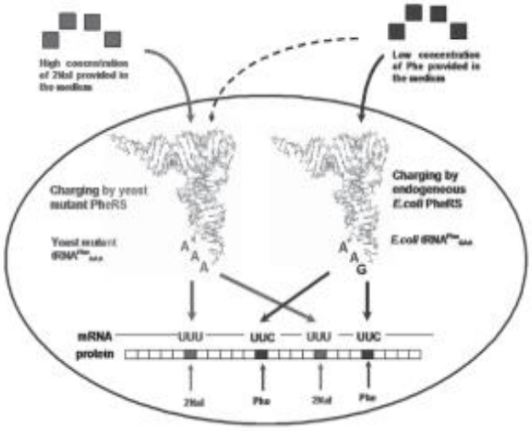


Fig. 2 Breaking the degeneracy of phenylalanine codons in *Escherichia coli*. The endogenous *E. coli* phenylalanyl-tRNA synthetase (*PheRS*) charges Phe to $tRNA^{Phe}_{GAA}$. The plasmid-borne yeast PheRS charges 2-naphthylalanine (*2Nal*) to yeast $tRNA^{Phe}_{AAA}$. UUC codons are decoded predominantly as Phe, while UUU codons are decoded predominantly as 2-naphthylalanine. *mRNA* messenger RNA, *tRNA* transfer RNA. (Reprinted with permission from Kwon et al. 2003. Copyright 2003 American Chemical Society)

Reassignment of sense codons can also be used for site-specific incorporation of noncanonical amino acids, although the fidelity of the method is lower than that of nonsense or frameshift suppression (Fig. 2). Because the 20 canonical amino acids are encoded by 61 sense codons, the genetic code is highly degenerate. For example, phenylalanine is coded by two codons, UUC and UUU. In E. coli, both codons are read by a single tRNA, which decodes UUC via Watson–Crick base-pairing and UUU through a “wobble” interaction. Reassignment of the UUU codon was achieved by introducing into an E. coli expression host a mutant yeast PheRS capable of charging 2-naphthylalanine, and a mutant yeast tRNAPhe equipped with an AAA anticodon. Expression of dehydrofolatereductase led to preferential incorporation of phenylalanine at UUC codons and of 2-naphthylalanine at UUU codons. The generality and quantitative specificity of this method have not yet been established.

Translational Fidelity

Aminoacyl-tRNASynthetases Translational fidelity is controlled in large measure by the aminoacyl-tRNAsynthetases, which match the 20 canonical amino acids with their cognate tRNAs. The remarkable capacity of the synthetases to discriminate among the natural amino acids might lead one to expect noncanonical substrates to be excluded by the translational apparatus (for more details see the chapter by Mascarenhas et al., this volume). In fact, many noncanonical amino acids are activated by the wild-type synthetases at rates that support efficient protein synthesis in bacterial cells. For analogs that are activated more slowly, addition of plasmid-encoded copies of the cognate synthetase can restore the rate of protein synthesis to levels characteristic of overexpressed recombinant proteins, and synthetase engineering has enabled further expansion of the set of useful amino acids. Szostak and coworkers have described a screen for identifying noncanonical amino acid substrates that are susceptible to enzymatic aminoacylation. Using the screen, they identified 59 previously unknown amino acid substrates.

Choice of protein scaffold for protein engineering

When engineering a new functionality in a protein, many aspects must be taken into account. First, it is necessary to know as much as possible about the starting structure in order to assess its potential. It is important to consider whether the protein will work in a particular selection or screening system, whether it will tolerate the changes introduced, and whether its production is simple and scalable enough for future applications. These are just a few examples of the assets to be considered in a structural framework.

The features we are seeking in a structural framework fall into two categories: experimental demands, and application demands. The fi rst is associated with the method used in the engineering experiment, while the latter depends on the intended use of the product.

A number of questions concerning the experimental procedure:

- Does an adequate assay, selection or screening system exist, or should the framework provide a means for testing the newly established functionality? If for example you are seeking a protein where signal change can be measured as a function of binding, certain scaffolds such as periplasmic binding proteins will facilitate this more easily than others
- Does the applied methodology have requirements for the framework? For example, small single - chain proteins are preferable for the application of phage and ribosome display and for the construction of fusion proteins. Likewise, cysteine - free scaffolds are useful when unique cysteines should be introduced to which effector compounds can be coupled. Further, a robust scaffold with high thermodynamic stability is preferable because it can compensate for any destabilizing effects of newly introduced functional residues, an effect often observed in rational design approaches.
- The expression of functional molecules is also important in selection and screening systems; for example, poorly or insolubly expressed proteins will not be able to complement a missing functionality and thus unstable variants are often eliminated in the process. A number of questions, concerning applicability:
- Under what conditions should the fi nal product be active, should it be especially stable or degradable, and does it need to be localized specifically?
- Is large - scale production feasible, what are the protein yields, and is there an easy purification?

- High thermodynamic stability, reversible folding, and high expression levels are what you will be looking for. The absence of disulfi de bonds or free cysteines is also advantageous because it allows the expression of functional molecules in the reducing environment of the bacterial cytoplasm, which usually produces higher yields than periplasmic or eukaryotic expression or refolding in vitro.

Proteins can be optimized to improve chemical robustness, thermodynamic stability or recombinant expression yields before using them as a framework in an engineering experiment. However, if considered well, the choice of a framework may also relieve the need to engineer many of these properties, so that attention can be focused on the property in question.

Apart from practical considerations, the choice of the structural framework can also be important for the new functionality that is introduced. Using a partial binding pocket and adjusting it to fi t a new ligand may be easier to achieve than introducing a new one from scratch. Obviously, studying the structure of the framework is essential in rational design approaches, but it can also be advantageous in directed evolution experiments. A detailed knowledge of the protein structure can reveal important parts that are better left untouched and help focus on the variable regions that can be subjected to randomization. To a certain degree, sequence alignments will also provide this type of information. Highly conserved residues are often important for folding or stability of the protein, while variable regions are free to evolve.

Applications of Molecular Modelling and Structure predictions to Protein engineering

Structure Predictions:

Protein structure prediction is the prediction of the three-dimensional structure of a protein from its amino acid sequence — that is, the prediction of its folding and its secondary, tertiary, and quaternary structure from its primary structure. Structure prediction is fundamentally different from the inverse problem of protein design. Protein structure prediction is one of the most important goals pursued by bioinformatics and theoretical chemistry; it is highly important in medicine (for example, in drug design) and biotechnology (for example, in the design of novel enzymes). Every two years, the performance of current methods is assessed in the CASP experiment (Critical Assessment of Techniques for Protein Structure Prediction). A continuous evaluation of protein structure prediction web servers is performed by the community project CAMEO3D.

Secondary structure prediction is a set of techniques in bioinformatics that aim to predict the local secondary structures of proteins based only on knowledge of their amino acid sequence only. For proteins, a prediction consists of assigning regions of the amino acid sequence as likely alpha helices, beta strands (often noted as "extended" conformations), or turns. The success of a prediction is determined by comparing it to the results of the DSSP algorithm (or similar e.g. STRIDE) applied to the crystal structure of the protein. Specialized algorithms have been developed for the detection of specific well-defined patterns such as transmembrane helices and coiled coils in proteins.

The best modern methods of secondary structure prediction in proteins reach about 80% accuracy; this high accuracy allows the use of the predictions as feature improving fold recognition and ab initio protein structure prediction, classification of structural motifs, and refinement of sequence alignments. The accuracy of current protein secondary structure prediction methods is assessed in weekly benchmarks such as LiveBench and EVA.

Background

Early methods of secondary structure prediction, introduced in the 1960s and early 1970s, focused on identifying likely alpha helices and were based mainly on helix-coil transition models. Significantly more accurate predictions that included beta sheets were introduced in the 1970s and relied on statistical assessments based on probability parameters derived from known solved structures. These methods, applied to a single sequence, are typically at most about 60-65% accurate, and often underpredict beta sheets.[1] The evolutionary conservation of secondary structures can be exploited by simultaneously assessing many homologous sequences in a multiple sequence alignment, by calculating the net secondary structure propensity of an aligned column of amino acids. In concert with larger databases of known protein structures and modern machine learning methods such as neural nets and support vector machines, these methods can achieve up 80% overall accuracy in globular proteins.[10] The theoretical upper limit of accuracy is around 90%,[10] partly due to idiosyncrasies in DSSP assignment near the ends of secondary structures, where local conformations vary under native conditions but may be forced to assume a single conformation in crystals due to packing constraints. Limitations are also imposed by secondary structure prediction's inability to account fortertiary structure; for example, a sequence predicted as a likely helix may still be able to adopt a beta-strand conformation if it is located within a beta-sheet region of the protein and its side chains pack well with their neighbors. Dramatic conformational changes related to the protein's function or environment can also alter local secondary structure.

Historical perspective

To date, over 20 different secondary structure prediction methods have been developed. One of the first algorithms was Chou-Fasman method, which relies predominantly on probability parameters determined from relative frequencies of each amino acid's appearance in each type of secondary structure. The original Chou-Fasman parameters, determined from the small sample of structures solved in the mid-1970s, produce poor results compared to modern methods, though the parameterization has been updated since it was first published. The Chou-Fasman method is roughly 50-60% accurate in predicting secondary structures.

The next notable program was the GOR method, named for the three scientists who developed it — *G*arnier, *O*sguthorpe, and *R*obson, is an information theory-based method. It uses the more powerful probabilistic technique of Bayesian inference. The GOR method takes into account not only the probability of each amino acid having a particular secondary structure, but also the conditional probability of the amino acid assuming each structure given the contributions of its neighbors (it does not assume that the neighbors have that same structure). The approach is both more sensitive and more accurate than that of Chou and Fasman because amino acid structural propensities are only strong for a small number of amino acids such as proline and glycine. Weak contributions from each of many neighbors can add up to strong effects overall. The original GOR method was roughly 65% accurate and is dramatically more successful in predicting alpha helices than beta sheets, which it frequently mispredicted as loops or disorganized regions.

Another big step forward, was using machine learning methods. First artificial neural networks methods were used. As a training sets they use solved structures to identify common sequence motifs associated with particular arrangements of secondary structures. These methods are over 70% accurate in their predictions, although beta strands are still often underpredicted due to the lack of three-dimensional structural information that would allow assessment of hydrogen bonding patterns that can promote formation of the extended conformation required for the presence of a complete beta sheet. PSIPRED and JPRED are some of the most known programs based on neural networks for protein

secondary structure prediction. Next, support vector machines have proven particularly useful for predicting the locations of turns, which are difficult to identify with statistical methods

Extensions of machine learning techniques attempt to predict more fine-grained local properties of proteins, such as backbone dihedral angles in unassigned regions. Both SVMs and neural networks have been applied to this problem. More recently, real-value torsion angles can be accurately predicted by SPINE-X and successfully employed for ab initio structure prediction.

Other improvements

It is reported that in addition to the protein sequence, secondary structure formation depends on other factors. For example, it is reported that secondary structure tendencies depend also on local environment,[18] solvent accessibility of residues, protein structural class,[20] and even the organism from which the proteins are obtained. Based on such observations, some studies have shown that secondary structure prediction can be improved by addition of information about protein structural class, residue accessible surface area and also contact number information.

Tertiary structure

The practical role of protein structure prediction is now more important than ever. Massive amounts of protein sequence data are produced by modern large-scale DNAsequencing efforts such as the Human Genome Project. Despite community-wide efforts in structural genomics, the output of experimentally determined protein structures—typically by time-consuming and relatively expensive X-ray crystallography or NMR spectroscopy—is lagging far behind the output of protein sequences.

The protein structure prediction remains an extremely difficult and unresolved undertaking. The two main problems are calculation of protein free energy and finding the global minimum of this energy. A protein structure prediction method must explore the space of possible protein structures which is astronomically large. These problems can be partially bypassed in "comparative" or homology modeling and fold recognition methods, in which the search space is pruned by the assumption that the protein in question adopts a structure that is close to the experimentally determined structure of another homologous protein. On the other hand, the *de novo* or ab initio protein structure predictionmethods must explicitly resolve these problems.

Ab initio protein modelling

Energy- and fragment-based methods

Ab initio- or *de novo*- protein modelling methods seek to build three-dimensional protein models "from scratch", i.e., based on physical principles rather than (directly) on previously solved structures. There are many possible procedures that either attempt to mimic protein folding or apply some stochastic method to search possible solutions (i.e.,global optimization of a suitable energy function). These procedures tend to require vast computational resources, and have thus only been carried out for tiny proteins. To predict protein structure *de novo* for larger proteins will require better algorithms and larger computational resources like those afforded by either powerful supercomputers (such as Blue Gene or MDGRAPE-3) or distributed computing (such as Folding@home, the Human Proteome Folding Project and Rosetta@Home). Although these computational barriers are vast, the potential benefits of

structural genomics (by predicted or experimental methods) make *ab initio* structure prediction an active research field.

As of 2009, a 50-residue protein could be simulated atom-by-atom on a supercomputer for 1 millisecond. As of 2012, comparable stable-state sampling could be done on a standard desktop with a new graphics card and more sophisticated algorithms.

Evolutionary covariation to predict 3D contacts

As sequencing became more commonplace in the 1990s several groups used protein sequence alignments to predict correlated mutations and it was hoped that these coevolved residues could be used to predict tertiary structure (using the analogy to distance constraints from experimental procedures such as NMR). The assumption is when single residue mutations are slightly deleterious, compensatory mutations may occur to restabilize residue-residue interactions. This early work used what are known as *local* methods to calculate correlated mutations from protein sequences, but suffered from indirect false correlations which result from treating each pair of residues as independent of all other pairs.

In 2011, a different, and this time *global* statistical approach, demonstrated that predicted coevolved residues were sufficient to predict the 3D fold of a protein, providing there are enough sequences available (>1,000 homologous sequences are needed). The method, EVfold, uses no homology modeling, threading or 3D structure fragments and can be run on a standard personal computer even for proteins with hundreds of residues. The accuracy of the contacts predicted using this and related approaches has now been demonstrated on many known structures and contact maps, including the prediction of experimentally unsolved transmembrane proteins.

Comparative protein modeling

Comparative protein modelling uses previously solved structures as starting points, or templates. This is effective because it appears that although the number of actual proteins is vast, there is a limited set of tertiary structural motifs to which most proteins belong. It has been suggested that there are only around 2,000 distinct protein folds in nature, though there are many millions of different proteins.

These methods may also be split into two groups:

Homology modeling

is based on the reasonable assumption that two homologous proteins will share very similar structures. Because a protein's fold is more evolutionarily conserved than its amino acid sequence, a target sequence can be modeled with reasonable accuracy on a very distantly related template, provided that the relationship between target and template can be discerned through sequence alignment. It has been suggested that the primary bottleneck in comparative modelling arises from difficulties in alignment rather than from errors in structure prediction given a known-good alignment. Unsurprisingly, homology modelling is most accurate when the target and template have similar sequences.

Protein threading

scans the amino acid sequence of an unknown structure against a database of solved structures. In each case, a scoring function is used to assess the compatibility of the sequence to the structure, thus yielding possible three-dimensional models. This type of method is also known as **3D-1D fold recognition** due to its compatibility analysis between three-dimensional structures and linear protein sequences. This method has also given rise to methods performing an **inverse folding search** by evaluating the compatibility of a given structure with a large database of sequences, thus predicting which sequences have the potential to produce a given fold.

Side-chain geometry prediction

Accurate packing of the amino acid side chains represents a separate problem in protein structure prediction. Methods that specifically address the problem of predicting side-chain geometry include dead-end elimination and the self-consistent mean field methods. The side chain conformations with low energy are usually determined on the rigid polypeptide backbone and using a set of discrete side chain conformations known as "rotamers." The methods attempt to identify the set of rotamers that minimize the model's overall energy.

These methods use rotamer libraries, which are collections of favorable conformations for each residue type in proteins. Rotamer libraries may contain information about the conformation, its frequency, and the standard deviations about mean dihedral angles, which can be used in sampling. Rotamer libraries are derived from structural bioinformatics or other statistical analysis of side-chain conformations in known experimental structures of proteins, such as by clustering the observed conformations for tetrahedral carbons near the staggered (60°, 180°, -60°) values.

Rotamer libraries can be backbone-independent, secondary-structure-dependent, or backbone-dependent. Backbone-independent rotamer libraries make no reference to backbone conformation, and are calculated from all available side chains of a certain type (for instance, the first example of a rotamer library, done by Ponder and Richards at Yale in 1987). Secondary-structure-dependent libraries present different dihedral angles and/or rotamer frequencies for α-helix, β-sheet, or coil secondary structures. Backbone-dependent rotamer libraries present conformations and/or frequencies dependent on the local backbone conformation as defined by the backbone dihedral angles ϕ and ψ, regardless of secondary structure.

The modern versions of these libraries as used in most software are presented as multidimensional distributions of probability or frequency, where the peaks correspond to the dihedral-angle conformations considered as individual rotamers in the lists. Some versions are based on very carefully curated data and are used primarily for structure validation, while others emphasize relative frequencies in much larger data sets and are the form used primarily for structure prediction, such as the Dunbrackrotamer libraries.

Side-chain packing methods are most useful for analyzing the protein's hydrophobic core, where side chains are more closely packed; they have more difficulty addressing the looser constraints and higher flexibility of surface residues, which often occupy multiple rotamer conformations rather than just one.

Prediction of structural classes

Statistical methods have been developed for predicting structural classes of proteins based on their amino acid composition, pseudo amino acid composition and functional domain composition.

Quaternary structure

In the case of complexes of two or more proteins, where the structures of the proteins are known or can be predicted with high accuracy, protein–protein docking methods can be used to predict the structure of the complex. Information of the effect of mutations at specific sites on the affinity of the complex helps to understand the complex structure and to guide docking methods.

Molecular modelling:

Molecular modelling encompasses all theoretical methods and computational techniques used to model or mimic the behaviour of molecules. The techniques are used in the fields of computational chemistry, drug design, computational biology and materials science for studying molecular systems ranging from small chemical systems to large biological molecules and material assemblies. The simplest calculations can be performed by hand, but inevitably computers are required to perform molecular modelling of any reasonably sized system. The common feature of molecular modelling techniques is the atomistic level description of the molecular systems. This may include treating atoms as the smallest individual unit (the Molecular mechanics approach), or explicitly modeling electrons of each atom (the quantum chemistry approach).

Molecular mechanics

Molecular mechanics is one aspect of molecular modelling, as it refers to the use of classical mechanics/Newtonian mechanics to describe the physical basis behind the models. Molecular models typically describe atoms (nucleus and electrons collectively) as point charges with an associated mass. The interactions between neighbouring atoms are described by spring-like interactions (representing chemical bonds) and van der Waals forces. The Lennard-Jones potential is commonly used to describe van der Waals forces. The electrostatic interactions are computed based on Coulomb's law. Atoms are assigned coordinates in Cartesian space or in internal coordinates, and can also be assigned velocities in dynamical simulations. The atomic velocities are related to the temperature of the system, a macroscopic quantity. The collective mathematical expression is known as a potential function and is related to the system internal energy (U), a thermodynamic quantity equal to the sum of potential and kinetic energies. Methods which minimize the potential energy are known as energy minimization techniques (e.g., steepest descent and conjugate gradient), while methods that model the behaviour of the system with propagation of time are known as molecular dynamics.

$$E = E_{\text{bonds}} + E_{\text{angle}} + E_{\text{dihedral}} + E_{\text{non-bonded}}$$
$$E_{\text{non-bonded}} = E_{\text{electrostatic}} + E_{\text{van der Waals}}$$

This function, referred to as a potential function, computes the molecular potential energy as a sum of energy terms that describe the deviation of bond lengths, bond angles and torsion angles away from equilibrium values, plus terms for non-bonded pairs of atoms describing van der Waals and electrostatic interactions. The set of parameters consisting of equilibrium bond lengths, bond angles, partial charge values, force constants and van der Waals parameters are collectively known as a force field. Different implementations of molecular mechanics use different mathematical expressions and different parameters for the potential function. The common force fields in use today have been developed by using high level quantum calculations and/or fitting to experimental data. The technique known as energy minimization is used to find positions of zero gradient for all atoms, in other words, a local energy minimum. Lower energy states are more stable and are commonly investigated because of their role in chemical and biological processes. A molecular dynamics simulation, on the other hand, computes the behaviour of a system as a function of time. It involves solving Newton's laws of motion, principally the second law, $\mathbf{F} = m\mathbf{a}$. Integration of Newton's laws of motion, using different integration algorithms, leads to atomic trajectories in

space and time. The force on an atom is defined as the negative gradient of the potential energy function. The energy minimization technique is useful for obtaining a static picture for comparing between states of similar systems, while molecular dynamics provides information about the dynamic processes with the intrinsic inclusion of temperature effects.

Variables

Molecules can be modelled either in vacuum or in the presence of a solvent such as water. Simulations of systems in vacuum are referred to as *gas-phase* simulations, while those that include the presence of solvent molecules are referred to as *explicit solvent* simulations. In another type of simulation, the effect of solvent is estimated using an empirical mathematical expression; these are known as *implicit solvation* simulations.

Applications

Molecular modelling methods are now routinely used to investigate the structure, dynamics, surface properties and thermodynamics of inorganic, biological and polymeric systems. The types of biological activity that have been investigated using molecular modelling include protein folding, enzymecatalysis, protein stability, conformational changes associated with biomolecular function, and molecular recognition of proteins, DNA, and membrane complexes.

Elementary introduction to Molecular Mechanics and Dynamics

Background

The "mechanical" molecular model was developed out of a need to describe molecular structures and properties in as practical a manner as possible. The range of applicability of molecular mechanics includes:

- Molecules containing thousands of atoms.
- Organics, oligonucleotides, peptides, and saccharides (metallo-organics and inorganics in some cases).
- Vacuum, implicit, or explicit solvent environments.
- Ground state only.
- Thermodynamic and kinetic (via molecular dynamics) properties.

The great computational speed of molecular mechanics allows for its use in procedures such as molecular dynamics, conformational energy searching, and docking. All the procedures require large numbers of energy evaluations.

Molecular mechanics methods are based on the following principles:

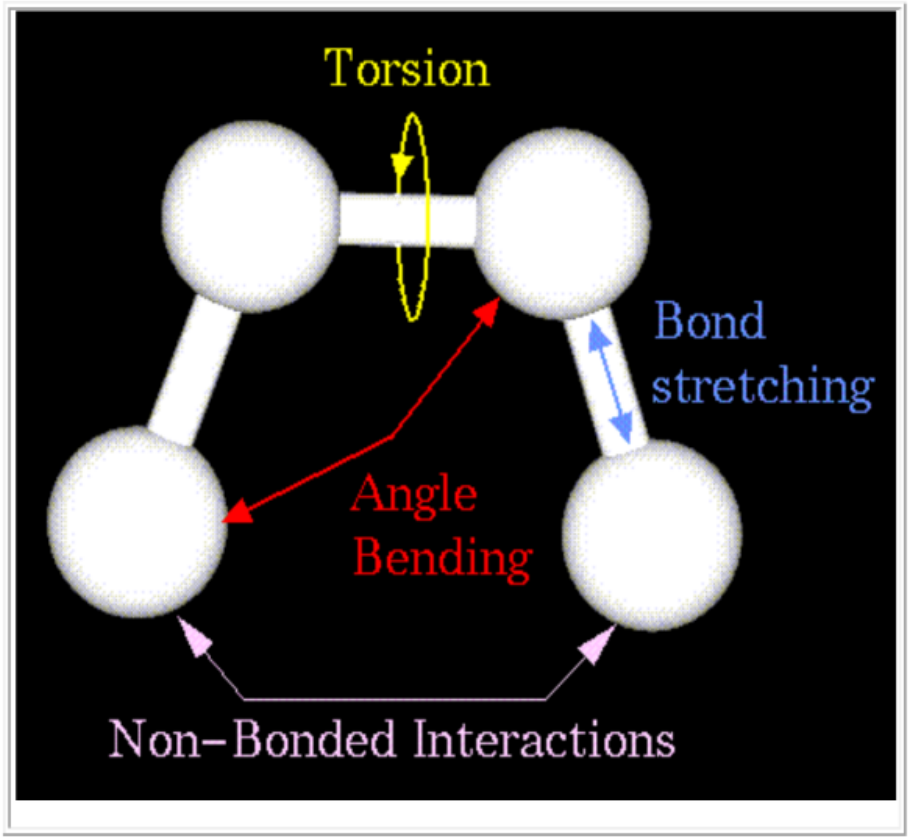
- Nuclei and electrons are lumped into atom-like particles.
- Atom-like particles are spherical (radii obtained from measurements or theory) and have a net charge (obtained from theory).
- Interactions are based on springs and classical potentials.
- Interactions must be preassigned to specific sets of atoms.

- Interactions determine the **spatial distribution** of atom-like particles and their **energies**.

Note how these principles differ from those of quantum mechanics.

The Anatomy of a Molecular Mechanics Force-Field

The mechanical molecular model considers atoms as spheres and bonds as springs. The mathematics of spring deformation can be used to describe the ability of bonds to stretch, bend, and twist:



Non-bonded atoms (greater than two bonds apart) interact through van der Waals attraction, steric repulsion, and electrostatic attraction/repulsion. These properties are easiest to describe mathematically when atoms are considered as spheres of characteristic radii.

The object of molecular mechanics is to predict the energy associated with a given conformation of a molecule. However, molecular mechanics energies have no meaning as absolute quantities. Only differences in energy between two or more conformations have meaning. A simple molecular mechanics energy equation is given by:

Energy = Stretching Energy +Bending Energy + Torsion Energy + Non-Bonded Interaction Energy

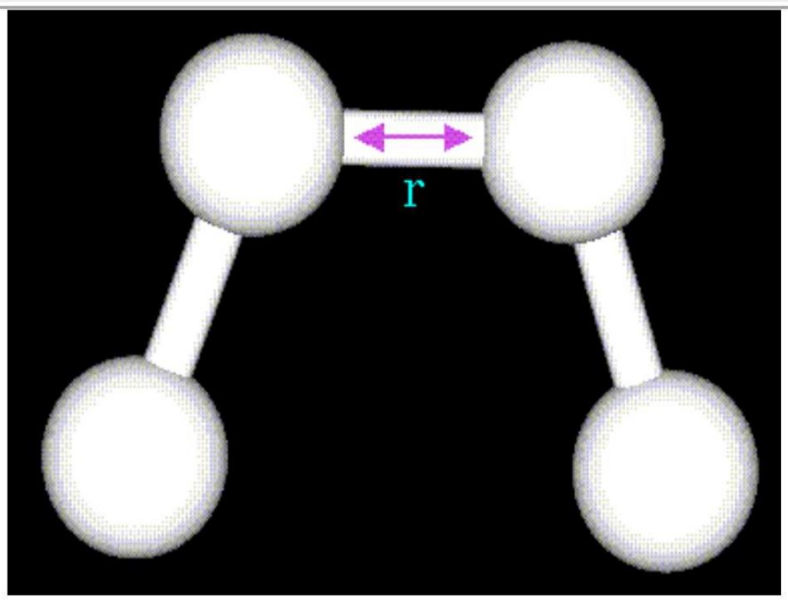
These equations together with the data (parameters) required to describe the behavior of different kinds of atoms and bonds, is called a force-field. Many different kinds of force-fields have been developed over the

years. Some include additional energy terms that describe other kinds of deformations. Some force-fields account for coupling between bending and stretching in adjacent bonds in order to improve the accuracy of the mechanical model.

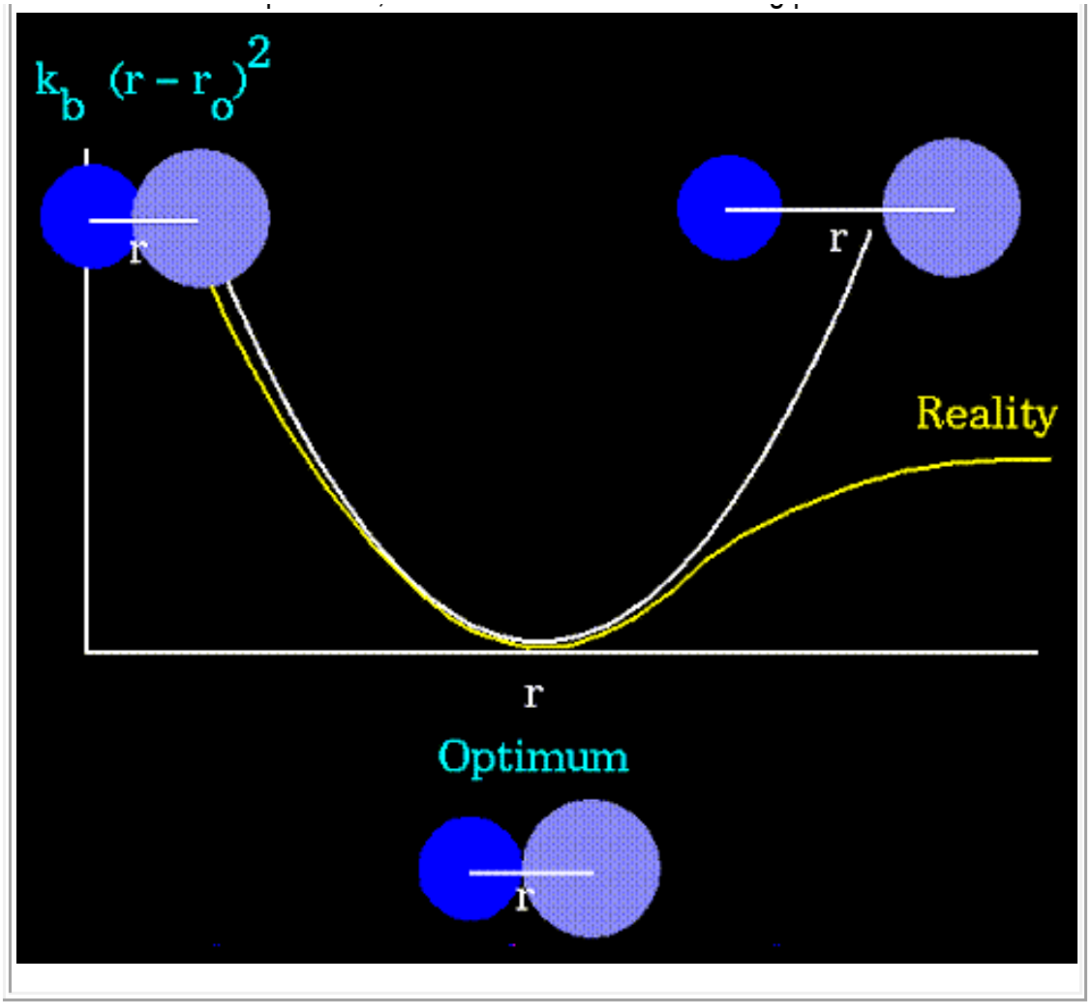
The mathematical form of the energy terms varies from force-field to force-field. The more common forms will be described.

- **Stretching Energy**

$$E = \sum_{bonds} k_b (r - r_o)^2$$

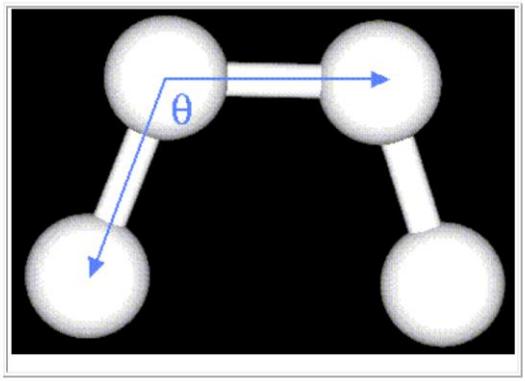


The stretching energy equation is based on Hooke's law. The "k*b*" parameter controls the stiffness of the bond spring, while "r*o*" defines its equilibrium length. Unique "k*b*" and "r*o*" parameters are assigned to each pair of bonded atoms based on their types (e.g. C-C, C-H, O-C, etc.). This equation estimates the energy associated with vibration about the equilibrium bond length. This is the equation of a parabola, as can be seen in the following plot:

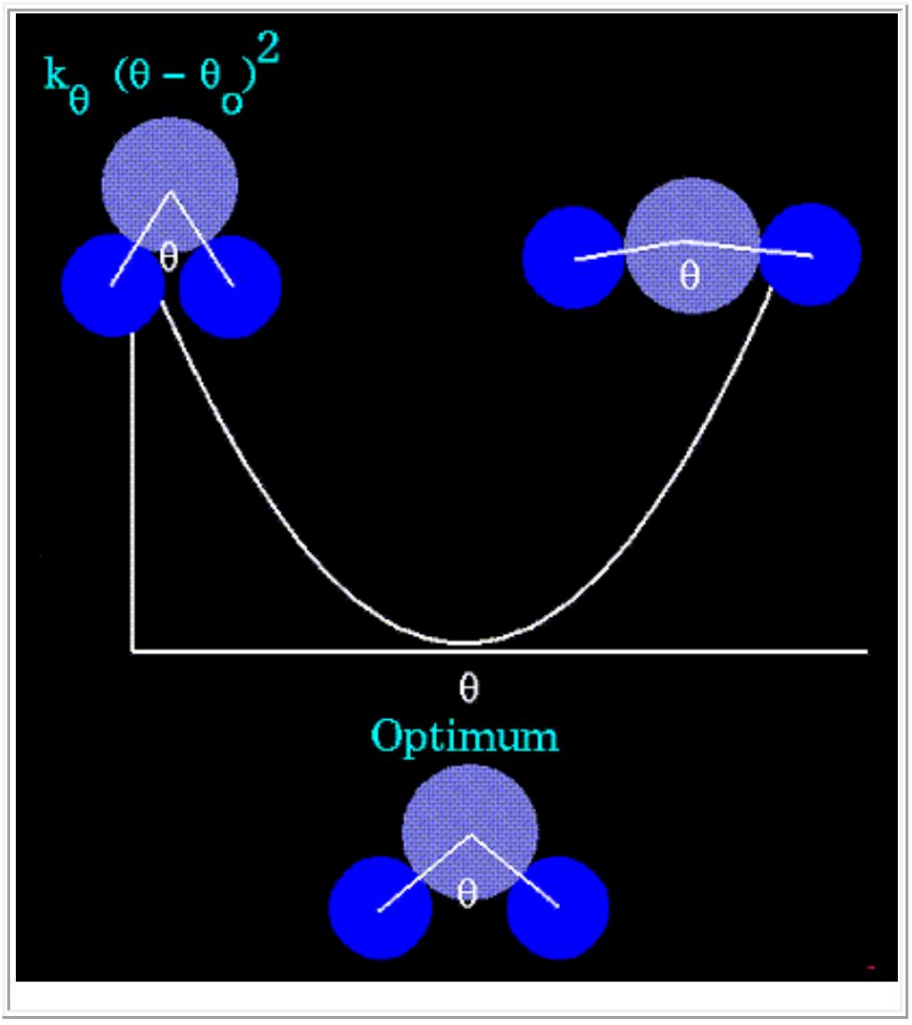


- Notice that the model tends to break down as a bond is stretched toward the point of dissociation.
- **Bending Energy**

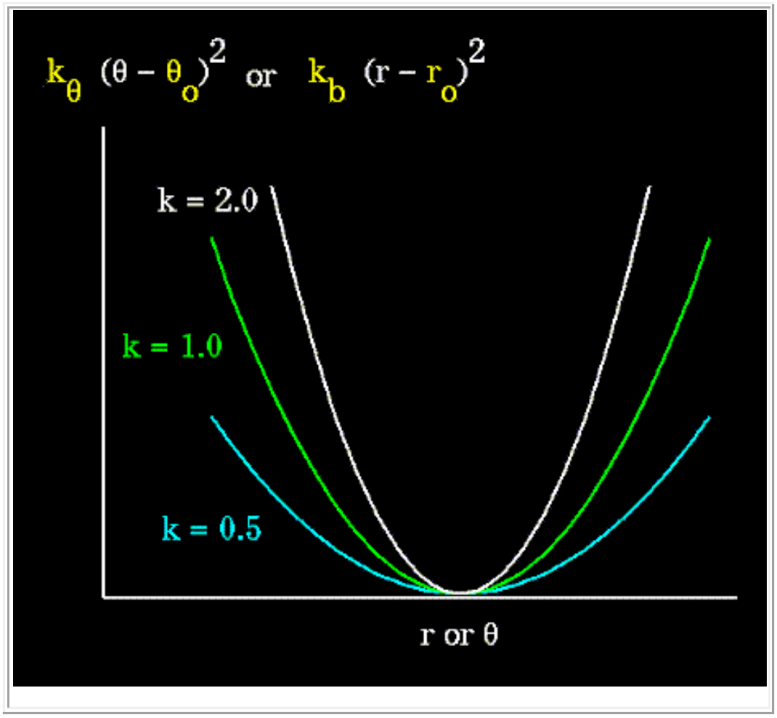
$$E = \sum_{\text{angles}} k_\theta (\theta - \theta_o)^2$$



- The bending energy equation is also based on Hooke's law. The "k*theta*" parameter controls the stiffness of the angle spring, while "theta*o*" defines its equilibrium angle. This equation estimates the energy associated with vibration about the equilibrium bond angle:

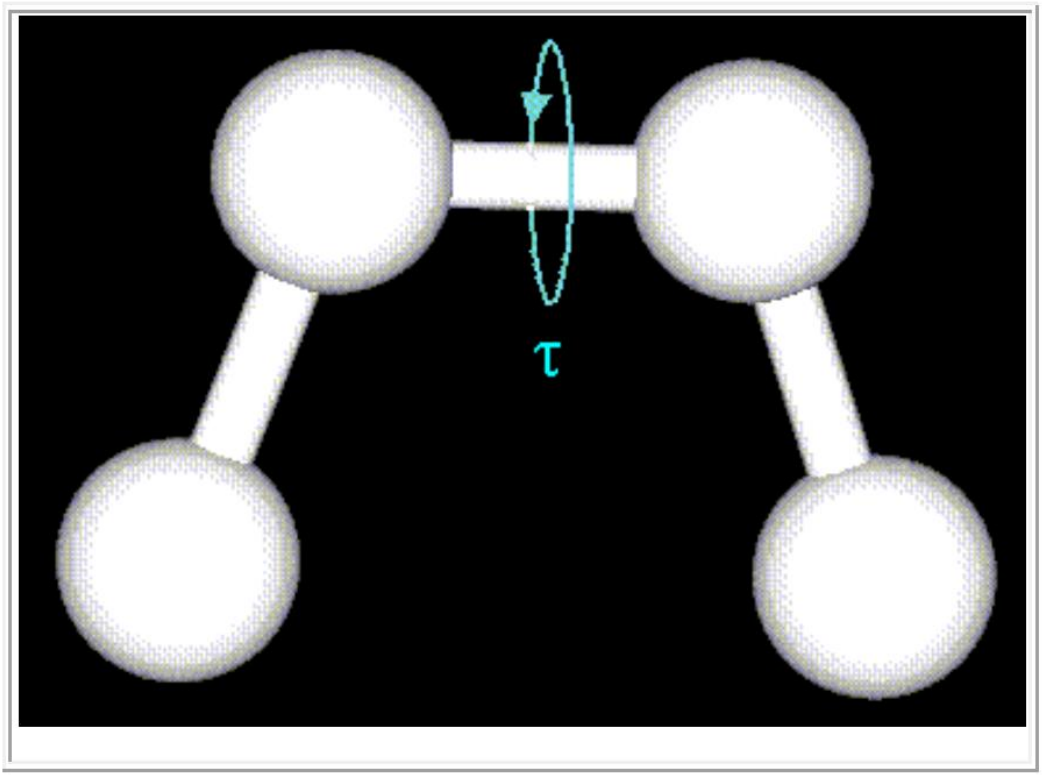


- Unique parameters for angle bending are assigned to each bonded triplet of atoms based on their types (e.g. C-C-C, C-O-C, C-C-H, etc.). The effect of the "k*b*" and "k*theta*" parameters is to broaden or steepen the slope of the parabola. The larger the value of "k", the more energy is required to deform an angle (or bond) from its equilibrium value. Shallow potentials are achieved for "k" values between 0.0 and 1.0. The Hookeian potential is shown in the following plot for three values of "k":

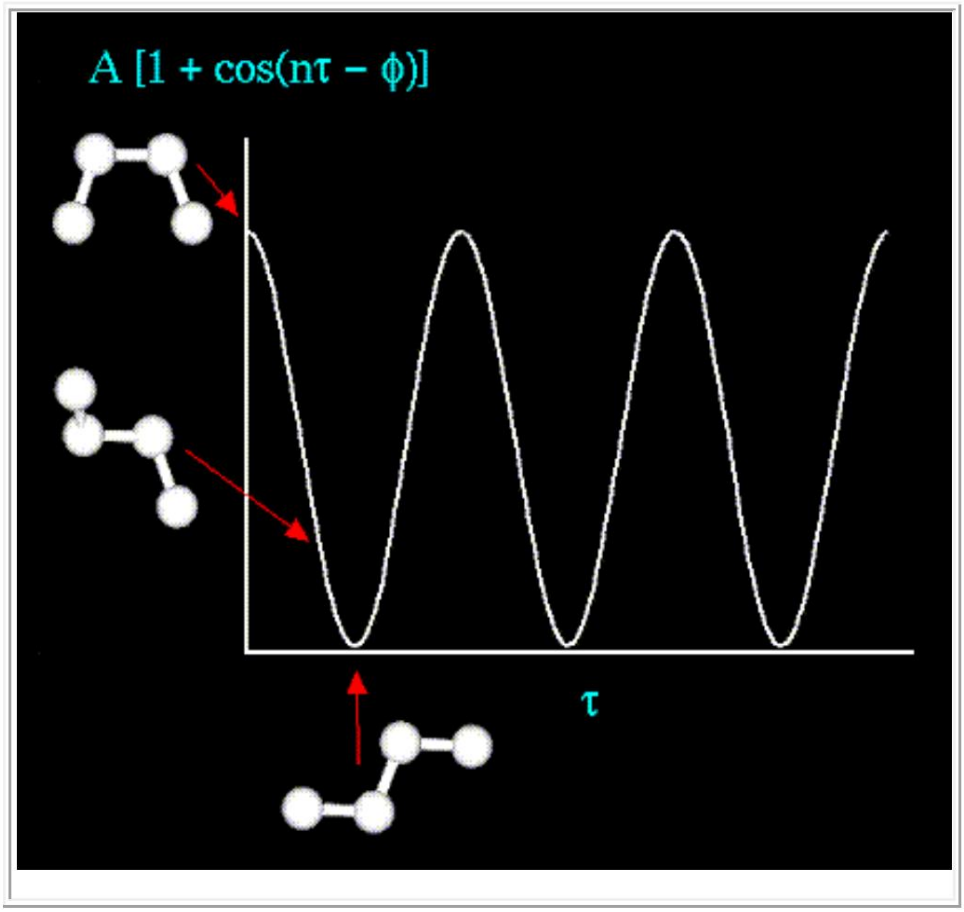


- **Torsion Energy**

$$E = \sum_{\text{torsions}} A\,[1 + \cos(n\tau - \phi)]$$

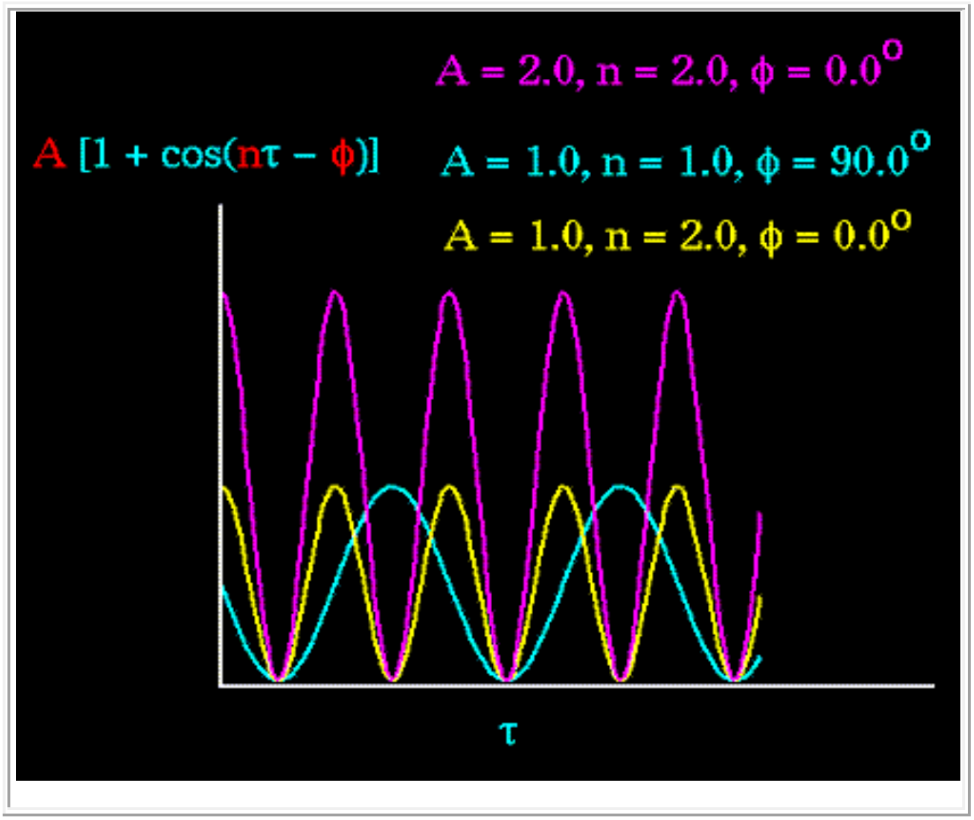


- The torsion energy is modeled by a simple periodic function, as can be seen in the following plot:



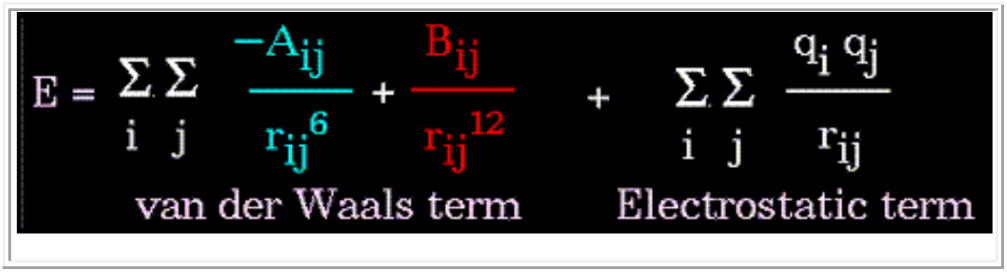
- The torsion energy in molecular mechanics is primarily used to correct the remaining energy terms rather than to represent a physical process. The torsional energy represents the amount of energy that must be added to or subtracted from the Stretching Energy + Bending Energy + Non-Bonded Interaction Energy terms to make the total energy agree with experiment or rigorous quantum

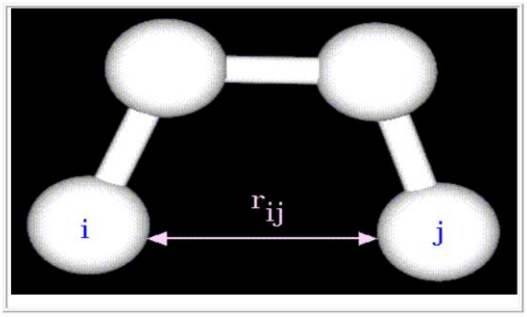
  mechanical calculation for a model dihedral angle (ethane, for example might be used aa model for any H-C-C-H bond).
- The "A" parameter controls the amplitude of the curve, the n parameter controls its periodicity, and "phi" shifts the entire curve along the rotation angle axis (tau). The parameters are determined from curve fitting. Unique parameters for torsional rotation are assigned to each bonded quartet of atoms based on their types (e.g. C-C-C-C, C-O-C-N, H-C-C-H, etc.). Torsion potentials with three combinations of "A", "n", and "phi" are shown in the following plot:



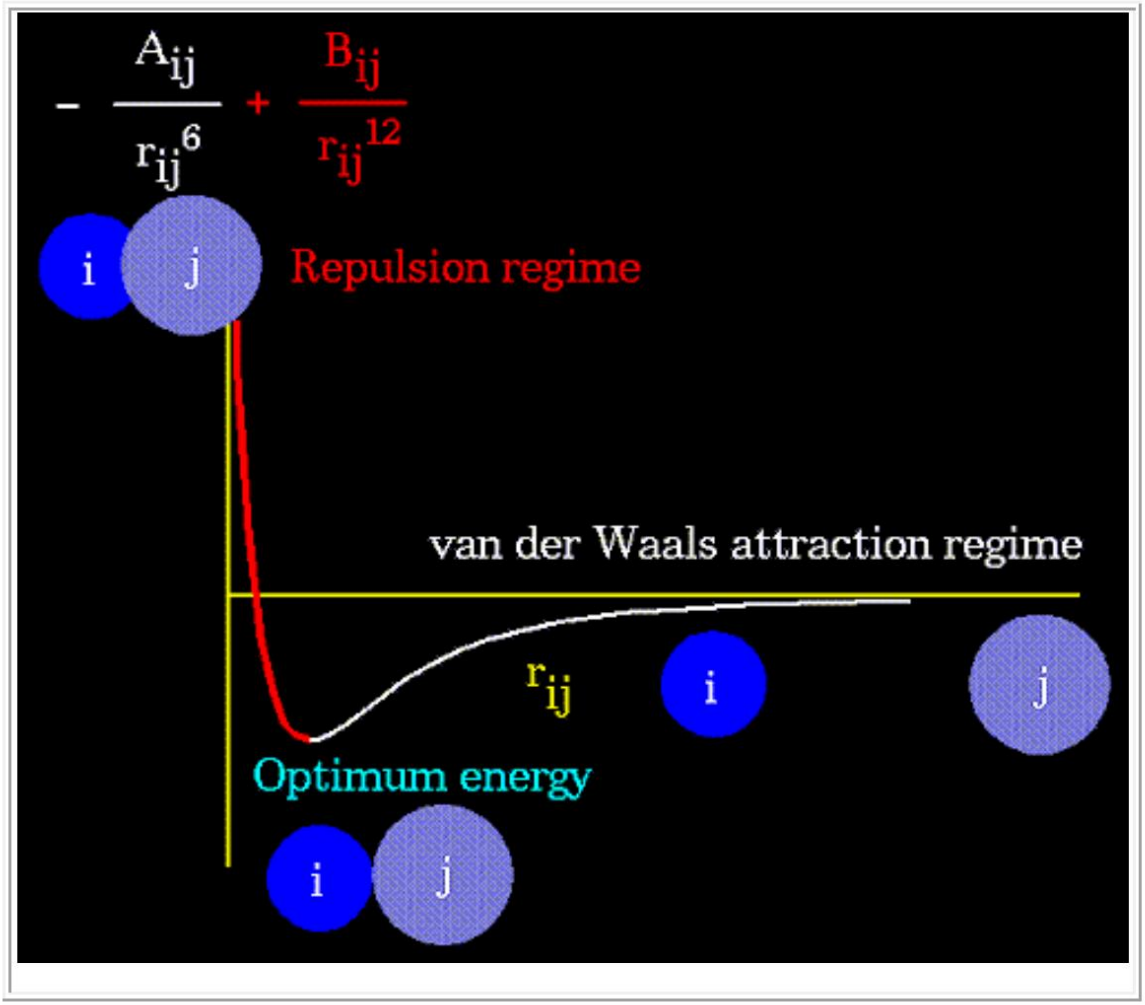
- Notice that "n" reflects the type symmetry in the dihedral angle. A CH3-CH3 bond, for example, ought to repeat its energy every 120 degrees. The *cis* conformation of a dihedral angle is assumed to be the zero torsional angle by convention. The parameter phi can be used to synchronize the torsional potential to the initial rotameric state of the molecule whose energy is being computed.
- **Non-Bonded Energy**

  The non-bonded energy represents the pair-wise sum of the energies of all possible interacting non-bonded atoms i and j:

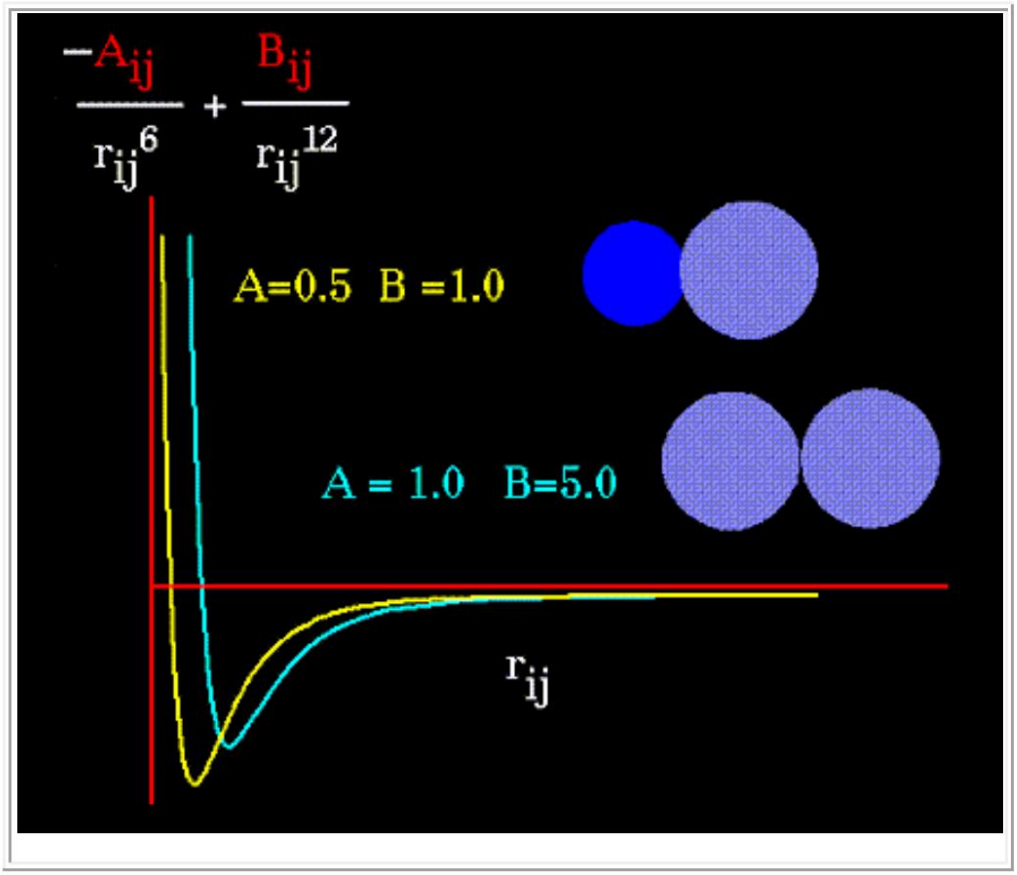




The non-bonded energy accounts for repulsion, van der Waals attraction, and electrostatic interactions. van der Waals attraction occurs at short range, and rapidly dies off as the interacting atoms move apart by a few Angstroms. Repulsion occurs when the distance between interacting atoms becomes even slightly less than the sum of their contact radii. Repulsion is modeled by an equation that is designed to rapidly blow up at close distances (1/r^12 dependency). The energy term that describes attraction/repulsion provides for a smooth transition between these two regimes. These effects are often modeled using a 6-12 equation, as shown in the following plot:



The "A" and "B" parameters control the depth and position (interatomic distance) of the potential energy well for a given pair of non-bonded interacting atoms (e.g. C:C, O:C, O:H, etc.). In effect, "A" determines the degree of "stickiness" of the van der Waals attraction and "B" determines the degree of "hardness" of the atoms (e.g marshmallow-like, billiard ball-like, etc.).



The "A" parameter can be obtained from atomic polarizability measurements, or it can be calculated quantum mechanically. The "B" parameter is typically derived from crystallographic data so as to reproduce observed average contact distances between different kinds of atoms in crystals of various molecules.

The electrostatic contribution is modeled using a Coulombic potential. The electrostatic energy is a function of the charge on the non-bonded atoms, their interatomic distance, and a molecular dielectric expression that accounts for the attenuation of electrostatic interaction by the environment (e.g. solvent or the molecule itself). Often, the molecular dielectric is set to a constant value between 1.0 and 5.0. A linearly varying distance-dependent dielectric (i.e. 1/r) is sometimes used to account for the increase in environmental bulk as the separation distance between interacting atoms increases.

Partial atomic charges can be calculated for small molecules using an *ab initio* or semiempirical quantum technique (usually MOPAC or AMPAC). Some programs assign charges using rules or templates, especially for macromolecules. In some force-fields, the torsional potential is calibrated to a particular charge calculation method (rarely made known to the user). Use of a different method can invalidate the force-field consistency.

References

1. Protein Engineering and Design 1st Edition, Edited By Sheldon J. Park, Jennifer R. Cochran; 2010

2. Enzyme Engineering and Evolution: Specific Enzyme Applications, Volume 644 - 1st Edition; Dan Tawfik. ISBN: 9780128244319

3. Enzyme Engineering by Preethi Kartan (Editor)

4. "Speeding Up the Protein Assembly Line". Genetic Engineering and Biotechnology News. 13 February 2015.

5. Wikipedia - https://en.wikipedia.org/wiki/Protein_structure