Computational Nanoscience

NSE C242 & Phys C203

Spring, 2008

Lecture 27:

Simulating Water and Examples in Computational Biology

April 29, 2008

Jeffrey C. Grossman

Elif Ertekin

Water

Water may seem to be quite an ordinary thing, made of two of the most reactive elements, O and H.

And yet, water is one of the most remarkable substances in the Universe.

It is the most studied material on Earth, and yet we still don't fully understand it!

If anything, in recent years there are more and more controversies regarding our understanding and the behavior of water. (e.g., recent Stanford X-ray experiments, NYT ice-skating article, etc.)

As one might imaging, there are many different ways to simulate the properties of water or take into account its impact on a system.

Today, we'll discuss some of these general approaches….but first, let's review why water is so cool.

(a lot of this stuff can be found at the excellent site: http://www.lsbu.ac.uk/water/

Water

Henry Cavendish was the first to describe correctly the composition of water (2 H + 1 O), in 1781.

He reported his findings in terms of phlogiston (later the gas he made was proven to be hydrogen) and dephlogisticated air (later this was proven to be oxygen).

Cavendish was a pretty neat guy.

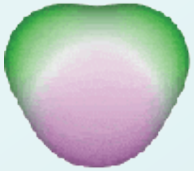
A University dropout, he also compared the electrical conductivities of equivalent solutions of electrolytes and expressed a version of Ohm's law.

H. Cavendish
THE HONOURABLE HENRY CAVENDISH.
Born 10th October 1731. Died 24th February 1810
(From a Drawing by Alexander in the Print Room of the British Museum)

His last major work was the first measurement of Sir Isaac Newton's gravitational constant, together with the mass and density of the Earth. The accuracy of this experiment was not improved on for nearly a century.

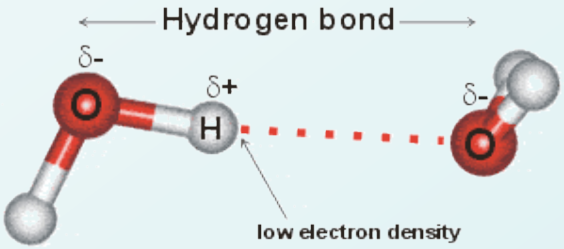
Water

Which one of these pictures best describes a water molecule?

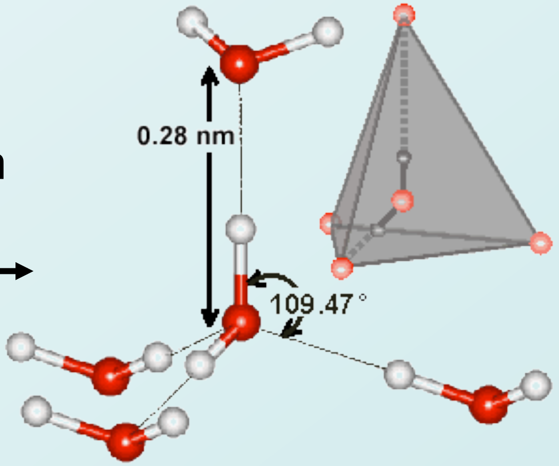


In fact, water is much more “rotund” than we typically draw - right hand picture is of actual charge density (pink negative, green positive).

The hydrogen bond, worth 0.1-0.2 eV is formed between two water molecules. Why is it preferred to be a straight line: O-H-O?



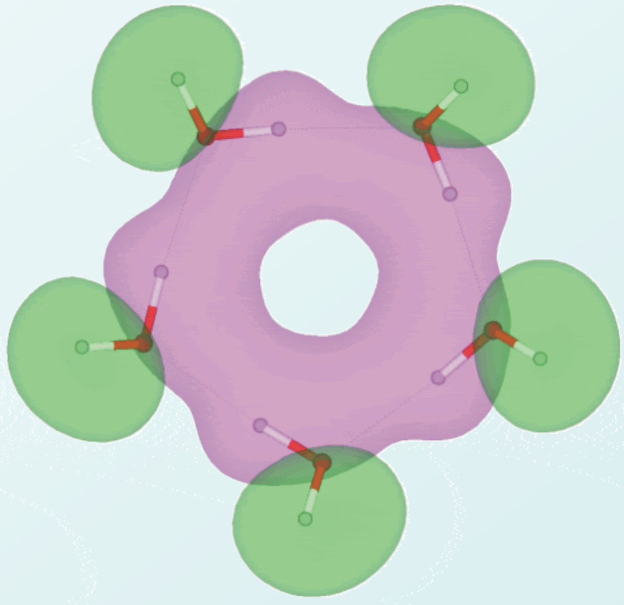
4 H bonds per molecules gives an ordered solid: ice



Water Pentamer

Despite such simplicity (i.e., a tetrahedral network) the structure of water is rather complex, particularly as temperature is increased.

For example, there 161 topographically distinct ways to put 5 water molecules together. Here's one of them:



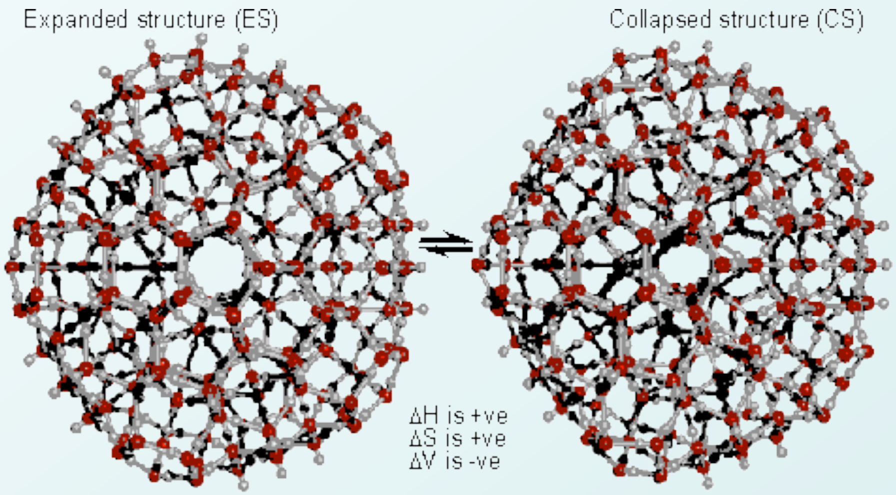
Still, the strong hydrogen bond combined with the preference for linearity has a very strong ordering effect as the temperature is lowered.

Water

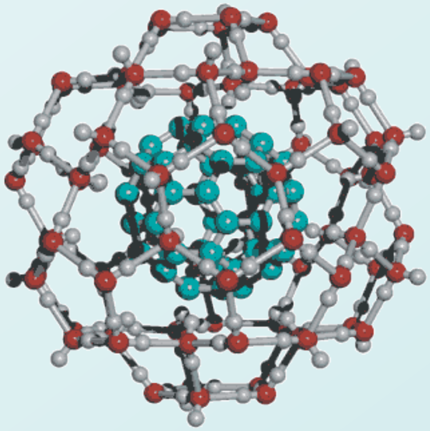
Water can form much more complex clusters.

Some research has postulated that water prefers to form icosahedra in varying sizes, and that this can explain some its properties.

For example, here’s a stable 280-water molecule cluster:



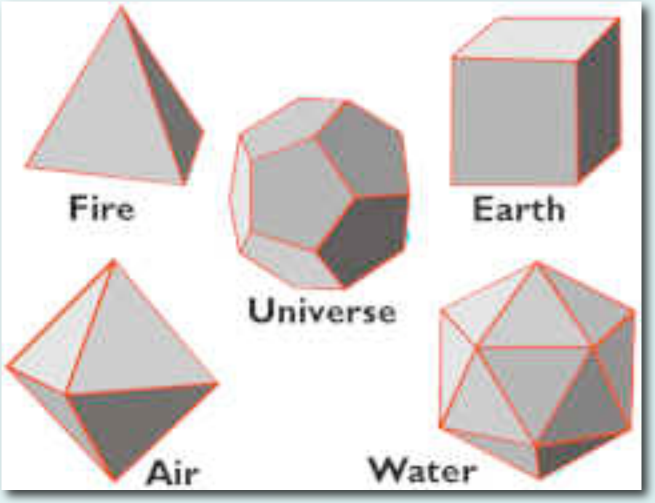
Proposed to solvate C_{60} →



Water

Here's a pretty cool connection from 2300 years ago:

Plato assumed 5 shapes "Platonic Solids" that had very distinct associations.



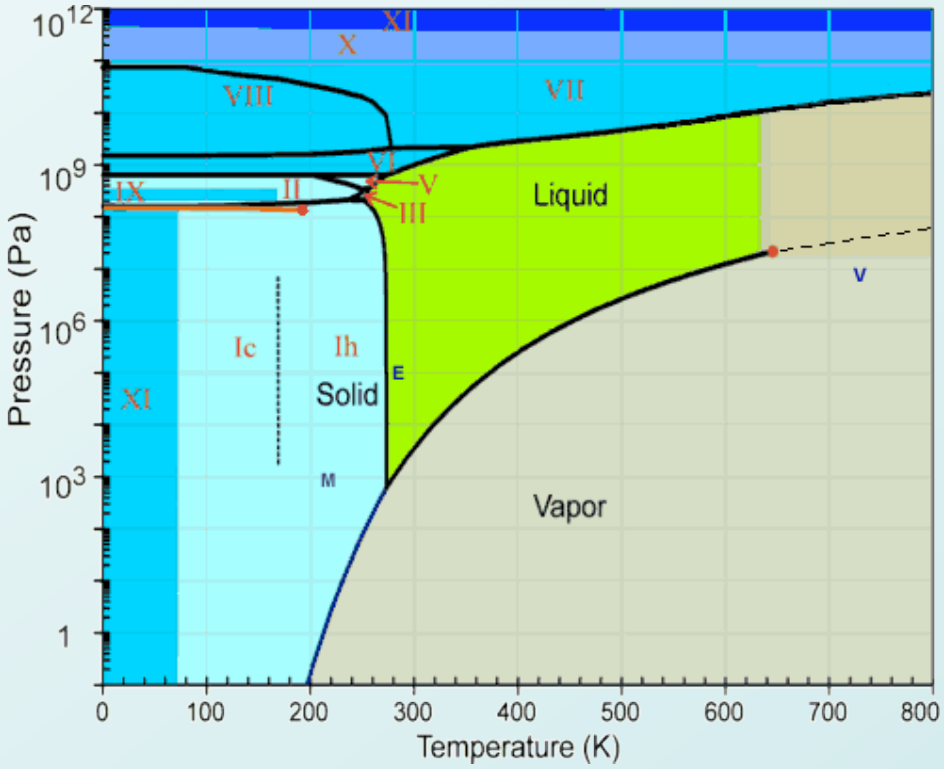
Water Phase Diagram

The phase diagram of water is complex, having a number of triple points and one or possibly two critical points.

Many of the crystalline forms may remain metastable in much of the low-temperature phase space at lower pressures.

There are 12-14 known phases of ice (so far), although many are still to be experimentally verified and are heavily debated.

Each one obeying the “ice rules” - 2 H atoms near each O atom and 1 H in between 2 O’s.

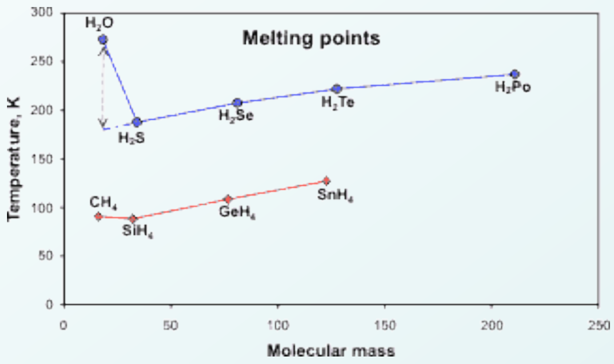


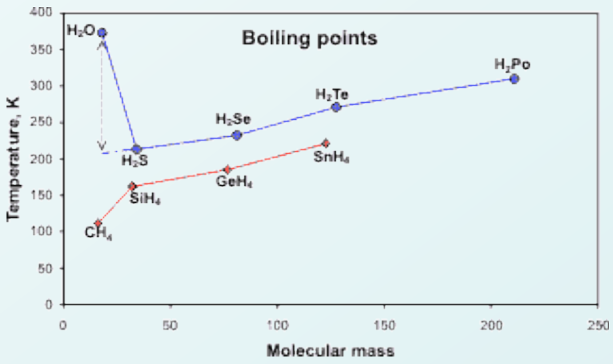
Jeffrey C. Grossman and Elif Ertekin, *NSE C242 & Phys C203, Spring 2008, U.C. Berkeley*

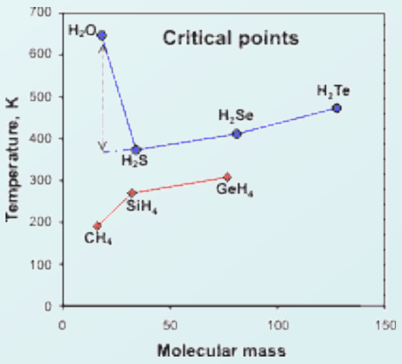
Water

There are many phenomena that make water so amazing.

Here are a few:







Of course, the most popular is probably that water shrinks upon melting, rather than expands (a usual 10% or so). Why is this?

How about “polywater” - do you all remember that?

Simulating Liquid Water

Let’s turn our attention to simulation.

We’ve already learned how to simulate a water molecule or a water cluster, so we will concentrate on the liquid.

There are effectively two types of water simulation: implicit or explicit.

A large number of models for water have been developed in order to discover the structure of water*.

The idea is that if “computer water” can successfully predict the physical properties of what we know about liquid water then the unknown structures and anomalies of liquid water could be determined.

Most water models involve orienting electrostatic effects and Lennard-Jones sites that may or may not coincide with one or more of the charged sites.

* See, for example, B.Guillot, “A reappraisal of what we have learnt during three decades of computer simulations on water,” J. Mol. Liquids **101** (2002) 219-260.

Simulating Liquid Water

The Lennard-Jones interaction accounts for the size of the molecules. It is repulsive at short distances, ensuring that the structure does not completely collapse due to the electrostatic interactions.

At intermediate distances LJ is significantly attractive but non-directional and competes with the directional attractive electrostatic interactions.

This competition ensures a tension between an expanded tetrahedral network and a collapsed non directional one (*e.g.* similar to that found in liquid noble gases).

Generally each model is developed to fit well with one particular physical structure or parameter (*e.g.* the density anomaly, radial distribution function or the critical parameters).

Some models are polarizable to make some allowance for charge redistribution and overall differences from one H2O to another; although most models are simpler and only try to reproduce “average” structures.

Explicit Water Models

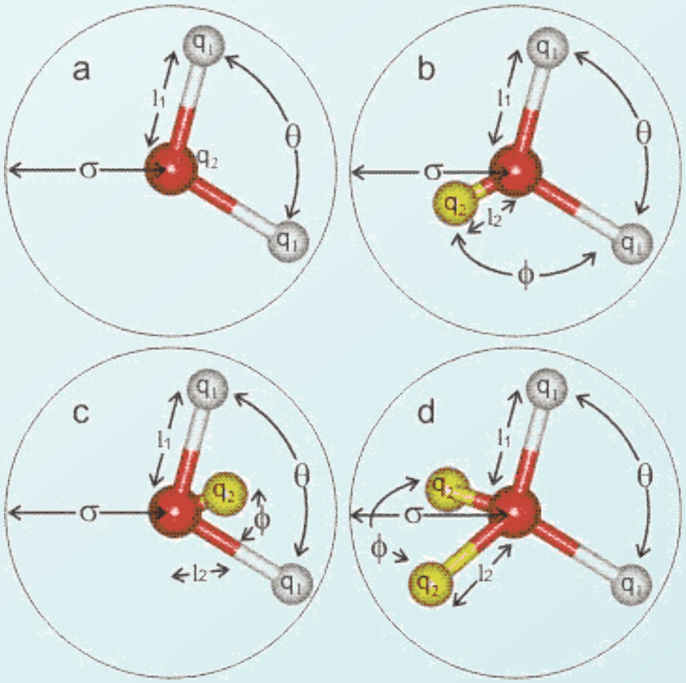
There are about 50 water models in the literature at this time.

Some of the aspects that vary include: rigid/flexible, fixed-charge/polarizable, number of charges, and how long-range the interaction can be.

Note, (a-c) are planar and (d) is tetrahedral

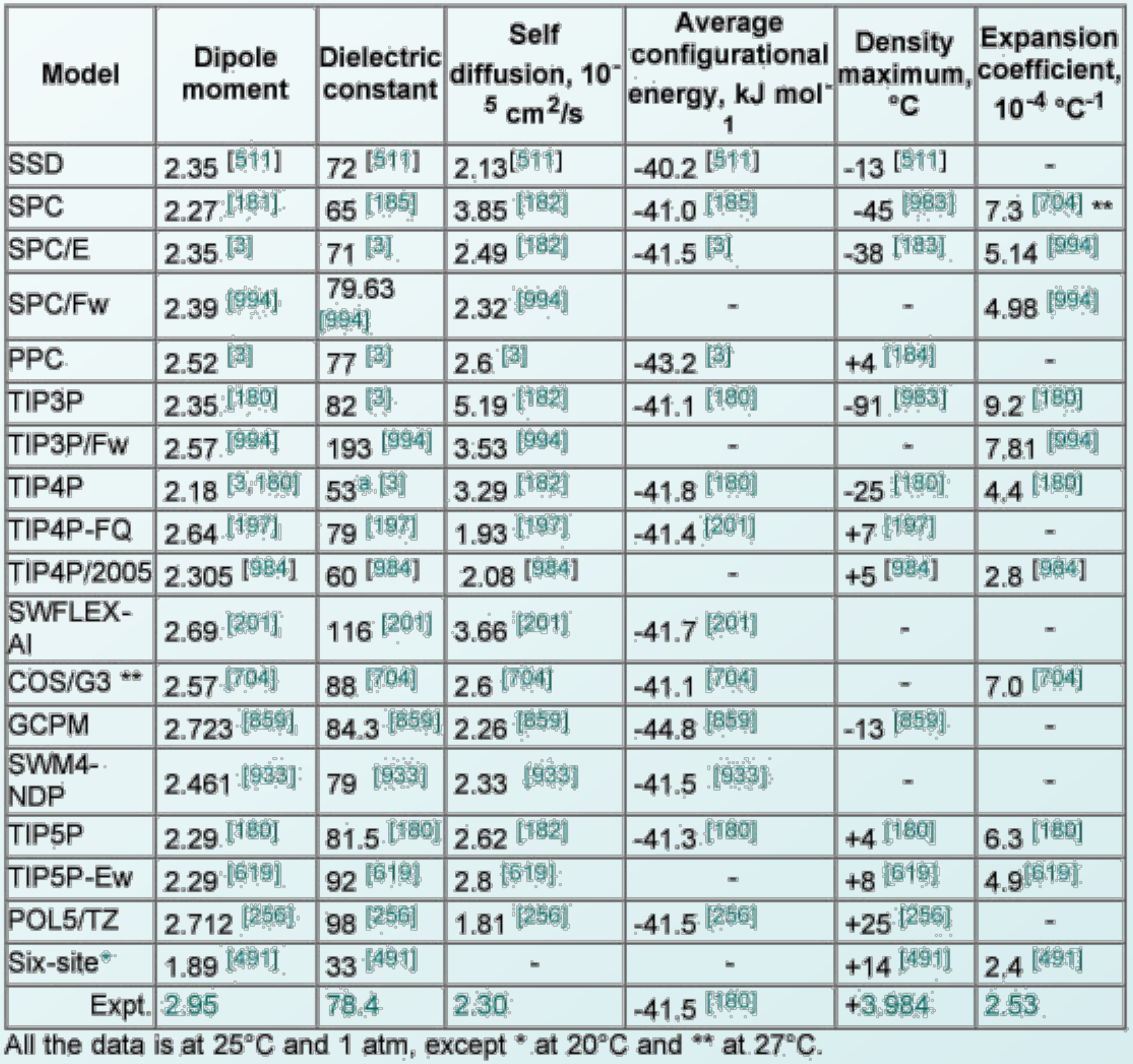
Some of popular ones are, for example,

(a) SPC, TIP3P
(b) PPC
(c) TIP4P, GPCM
(d) TIP5P, ST2



Although such simple models are of great utility, no universally applicable model can be identified at this time.

Some Data from Classical Models



Model	Dipole moment	Dielectric constant	Self diffusion, 10^{-5} cm^2/s	Average configurational energy, kJ mol^{-1}	Density maximum, °C	Expansion coefficient, 10^{-4} $°C^{-1}$
SSD	2.35 [511]	72 [511]	2.13[511]	-40.2 [511]	-13 [511]	-
SPC	2.27 [181]	65 [185]	3.85 [182]	-41.0 [185]	-45 [983]	7.3 [704] **
SPC/E	2.35 [3]	71 [3]	2.49 [182]	-41.5 [3]	-38 [183]	5.14 [994]
SPC/Fw	2.39 [994]	79.63 [994]	2.32 [994]	-	-	4.98 [994]
PPC	2.52 [3]	77 [3]	2.6 [3]	-43.2 [3]	+4 [184]	-
TIP3P	2.35 [180]	82 [3]	5.19 [182]	-41.1 [180]	-91 [983]	9.2 [180]
TIP3P/Fw	2.57 [994]	193 [994]	3.53 [994]	-	-	7.81 [994]
TIP4P	2.18 [3, 180]	53[a] [3]	3.29 [182]	-41.8 [180]	-25 [180]	4.4 [180]
TIP4P-FQ	2.64 [197]	79 [197]	1.93 [197]	-41.4 [201]	+7 [197]	-
TIP4P/2005	2.305 [984]	60 [984]	2.08 [984]	-	+5 [984]	2.8 [984]
SWFLEX-AI	2.69 [201]	116 [201]	3.66 [201]	-41.7 [201]	-	-
COS/G3 **	2.57 [704]	88 [704]	2.6 [704]	-41.1 [704]	-	7.0 [704]
GCPM	2.723 [859]	84.3 [859]	2.26 [859]	-44.8 [859]	-13 [859]	-
SWM4-NDP	2.461 [933]	79 [933]	2.33 [933]	-41.5 [933]	-	-
TIP5P	2.29 [180]	81.5 [180]	2.62 [182]	-41.3 [180]	+4 [180]	6.3 [180]
TIP5P-Ew	2.29 [619]	92 [619]	2.8 [619]	-	+8 [619]	4.9[619]
POL5/TZ	2.712 [256]	98 [256]	1.81 [256]	-41.5 [256]	+25 [256]	-
Six-site*	1.89 [491]	33 [491]	-	-	+14 [491]	2.4 [491]
Expt.	2.95	78.4	2.30	-41.5 [180]	+3.984	2.53

All the data is at 25°C and 1 atm, except * at 20°C and ** at 27°C.

Classical Water Models

So…assuming we are going to go with an explicit classical model for water, which one of those 50 should I use?

The answer, as is the case for all classical or semi-empirical simulation techniques, is that it depends.

If one is studying the impact of water on the dynamical aspects of a non-reactive process, then a simple model such as TIP3P or SPC may be just fine.

On the other hand, if one is investigating the hydrophobic interface where subtle charge redistributions can play a huge role, a polarizable model such as TIP4F may be necessary.

In general, as usual, it is strongly advised to test the potential for the specific property of interest before committed major (human and computer) resources.

Beyond Classical Simulations for Explicit Water

In some cases, one needs to have a better description of water than is possible from classical models.

Here, we must move to a quantum mechanical description for the electrons.

As we have seen, this is much slower in terms of computational speed, so we will be severely limited in terms of how big a system and for how long we can simulate.

For example,

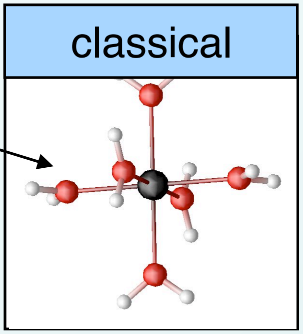
typical classical simulation:	10,000 water molecules, 1-10 ns
typical quantum simulation:	64 water molecules, 1-20 ps

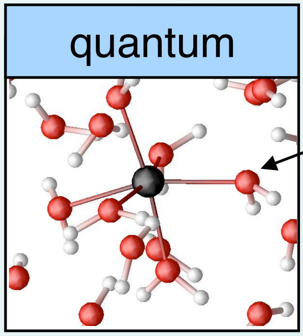
And yet, even with such limitations, quantum simulations of water have proven very important, both in substantiating which classical model to use as well as making genuine predictions comparable with experiment.

When Classical Models Are Not Good Enough

Here is a very simple example: a single Mg++ atom in water, where classical and quantum simulations show important differences*

tilt not observed in classical simulations

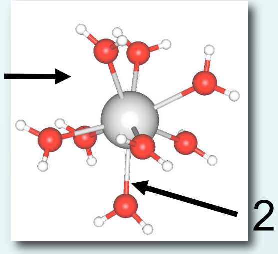




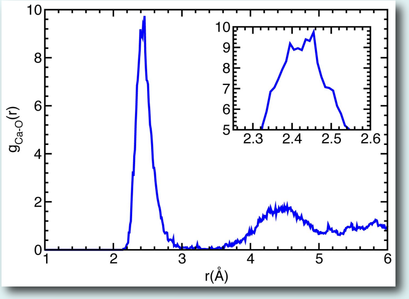
Ab initio simulations show lone-pair stretch is responsible

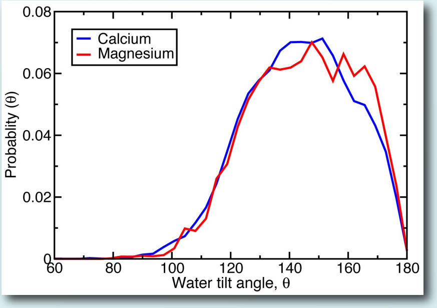
Subsequent experiments** agree with quantum simulations of Ca^{2+}

45º Tilt



2.43 Å Ca-O





The quantum calculations of g(r) and tilt angles are in excellent agreement with experiments

*Schwegler, Lightstone, Galli

**EXAFS and NDIS data,Fultan et al, PNL

Implicit Water

What if we still want a quantum mechanical picture for a solute in water, but not necessarily for the water itself?

This can be done (and is in fact very common) by introducing water as a simple “external potential.”

The advantage is, of course, that it is extremely fast and can give a rough estimate of the impact of a solvent.

The disadvantage is that there are no more water molecules in the system, so molecular-level changes or responses are averaged over and one cannot estimate, e.g., dynamical processes or distributions.

Still, these implicit water molecules are not to be undervalued!

Continuum Models

One of the most common implicit water models is the Polarizable Continuum Model, or PCM*.

In PCM, a solute molecule is treated quantum mechanically, and is placed in a volume, the “solute cavity”.

This cavity is a function of the molecular structure; in PCM, it is defined by a set of spheres centered on the nuclei of the solute molecule.

The cavity is surrounded by a continuum dielectric.

The solute molecule polarizes the dielectric; the dielectric polarization, in turn, generates an electrostatic field at the solute molecule, which modifies its electron density.

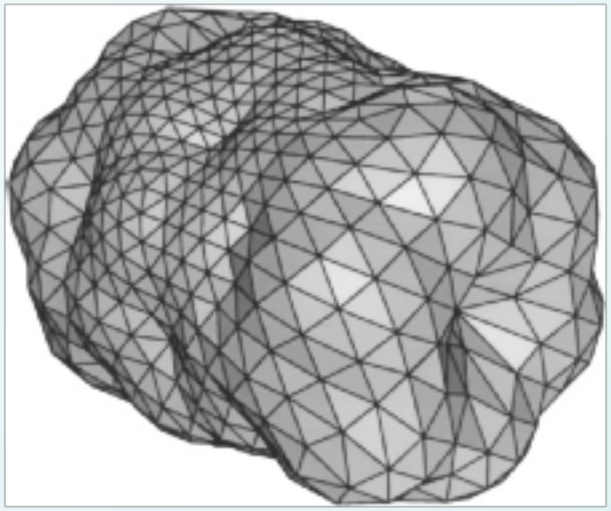
The solute-molecule - solvent interaction is expressed in terms of the interaction of the electrostatic potential of the solute molecule with some charge density on the surface of the cavity.

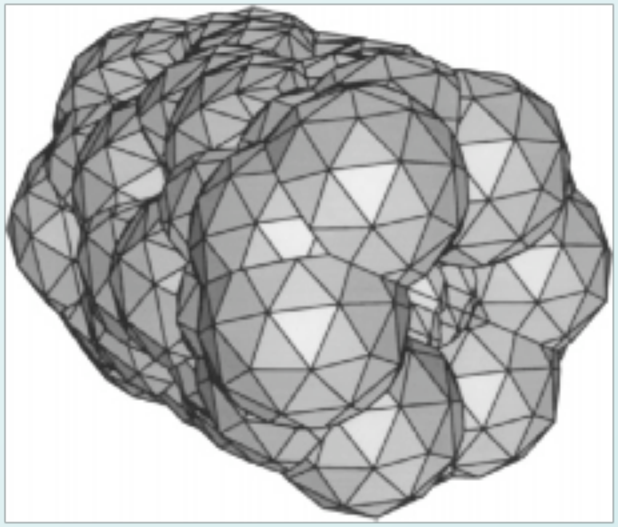
* Miertus, S.; Scrocco, E.; Tomasi, J. Chem. Phys. 1981, 55, 117.

Cavity is Key

The choice of how to define the solute cavity plays huge role in the accuracy of the continuum solvation model.

It is still a topic of intense research; for example here are two cavities for solvating a nanotube*:





* Journal of Computational Chemistry, Vol. 20, No. 16, 1693–1701 (1999)

An Abrupt Transition: Proteins

Proteins are polymers constructud from sequences of amino acids.

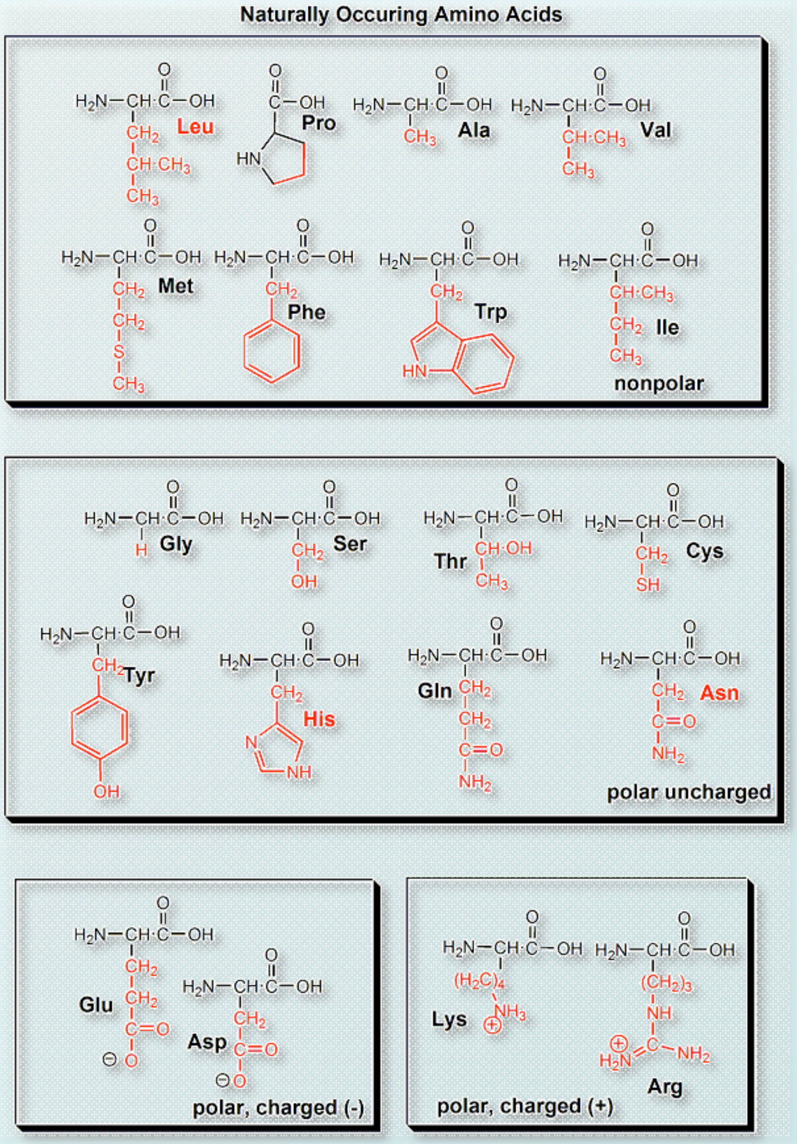
There are 20 common amino acids.

These are linked together by “amide” bonds to give a “polypeptide” chain.

By look at the variation among these simple molecules, one can see how much opportunity there is for different sizes, shapes, hydrogen bonding, and charge distributions.

This is why proteins can perform so many different biological functions…

…and also why they’re so difficult to simulate.



http://employees.csbsju.edu/hjakubowski/classes/ch112/proteins/aminoacidprot1.gif

Proteins

The biological function of a protein is intimately dependent on the conformation that the molecule can take.

In contrast to most synthetic polymers, a protein usual exists in only 1 ground state structure.

These structures ("native states") are what are found in typical living cell conditions (neutral pH at around 20-40 C).

Proteins can be unfolded ("denatured") using temperature or certain solvents.

The unfolding is reversible and so the protein can be folded back to its native structure.

And what might those structures be?? How do we determine them? Why do we care about them?

Protein Structure

X-ray crystallography and NMR are the two most widely used methods for determining a protein’s structure.

However, while new protein sequences are determined very quickly, the structure of a protein is much more difficult to ascertain and is therefore a much slower process.

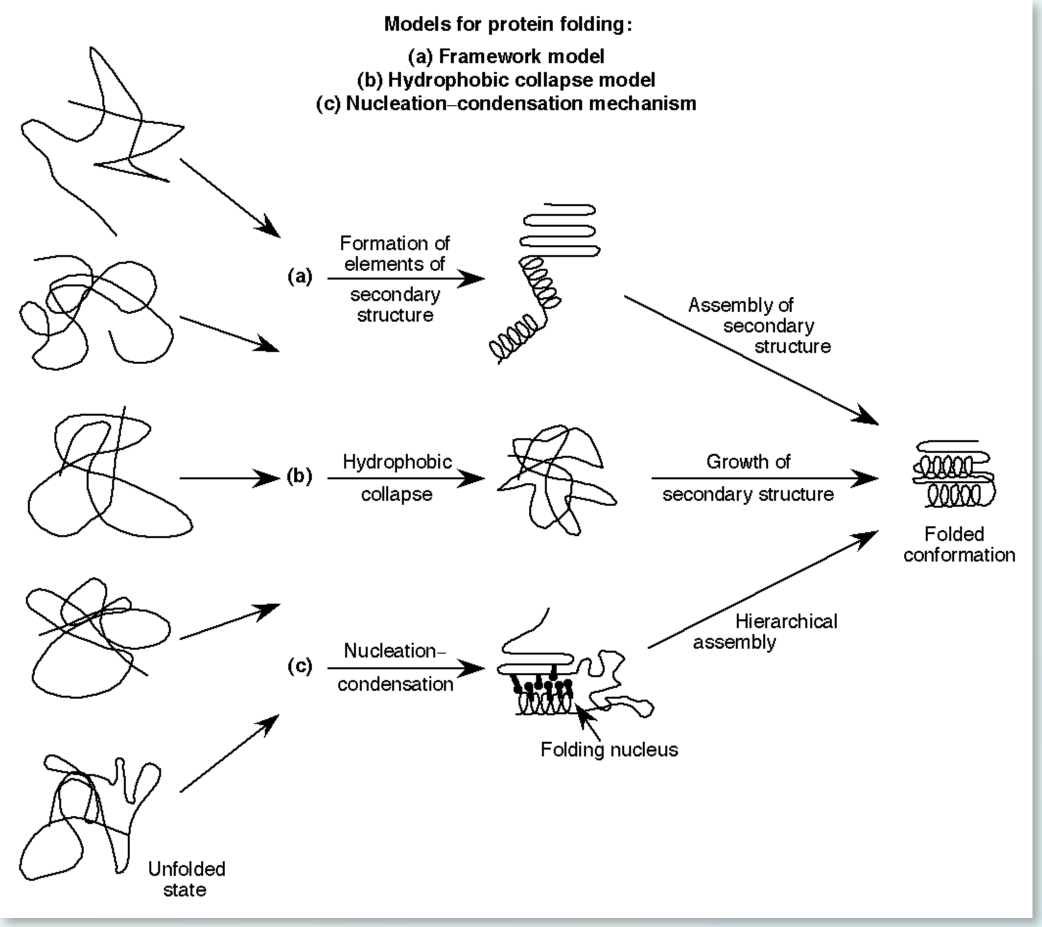
It is possible to some extent to assign function just from an analysis of the sequence alone.

However, the much stronger relationship between the 3-D structure and function of a protein makes functional assignment based on structure much more appealing.

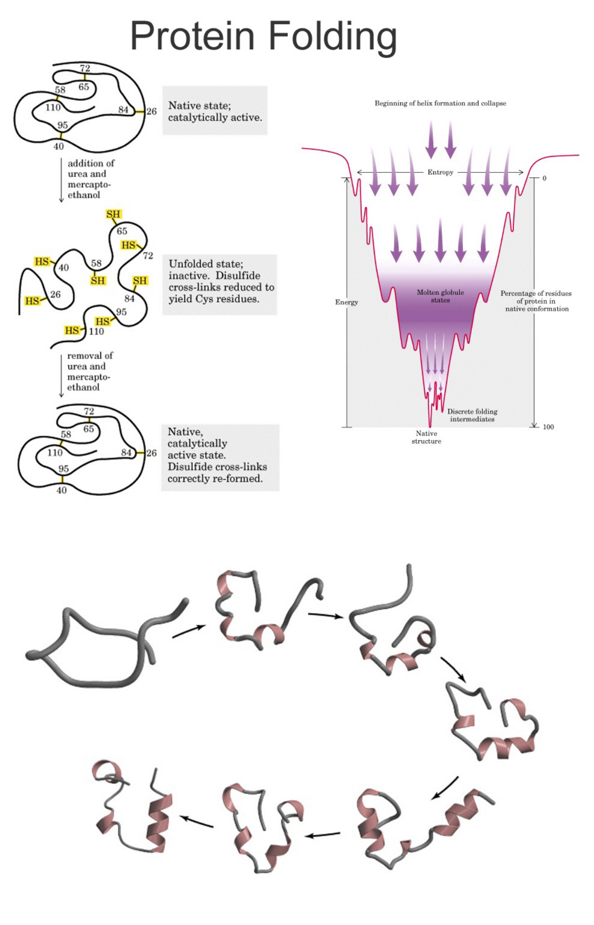
The fact that it is difficult to obtain a protein structure experimentally means that there is considerable interest in the application of theoretical methods for predicting the 3-D structure of proteins (given the amino acid sequence).

Protein Folding

The structural determination of a protein is sometimes generally referred to as the “protein folding problem”. As we can see, it looks pretty easy…right?



Models for protein folding:
(a) Framework model
(b) Hydrophobic collapse model
(c) Nucleation–condensation mechanism



Protein Folding

Native state; catalytically active.

Unfolded state; inactive. Disulfide cross-links reduced to yield Cys residues.

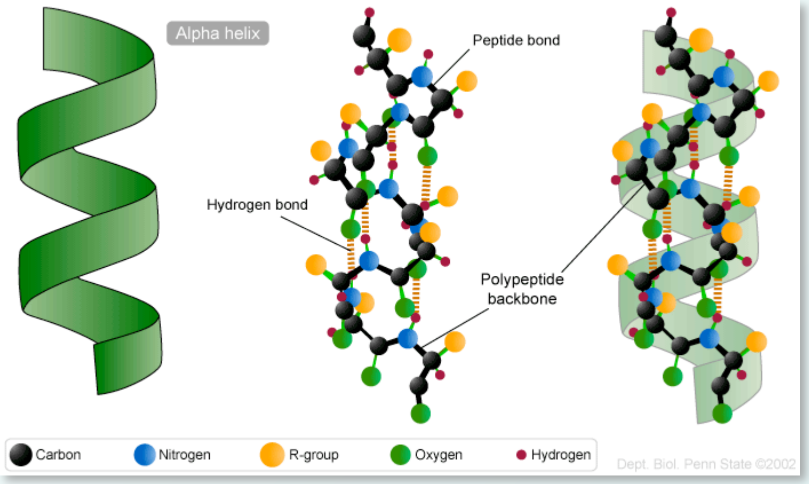
Native, catalytically active state. Disulfide cross-links correctly re-formed.

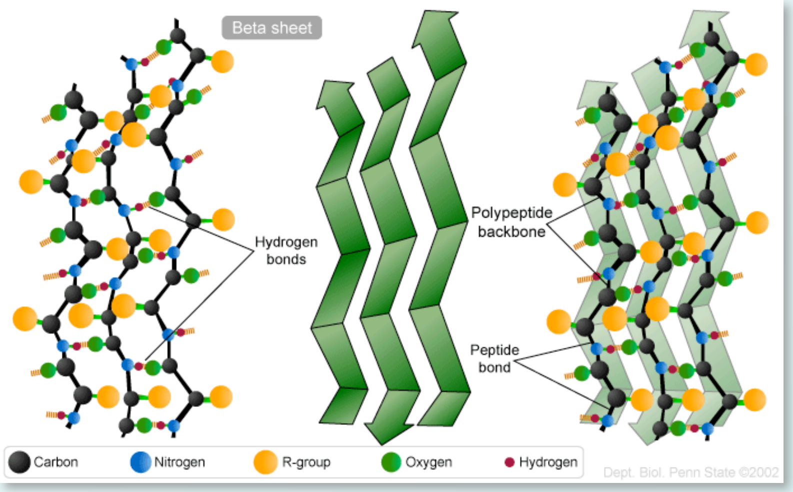
72
65
58
110
84
26
95
40
addition of urea and mercapto-ethanol
SH
HS
removal of urea and mercapto-ethanol
Beginning of helix formation and collapse
Entropy
0
Energy
Molten globule states
Percentage of residues of protein in native conformation
Discrete folding intermediates
Native structure
100

Jeffrey C. Grossman and Elif Ertekin, *NSE C242 & Phys C203, Spring 2008, U.C. Berkeley*

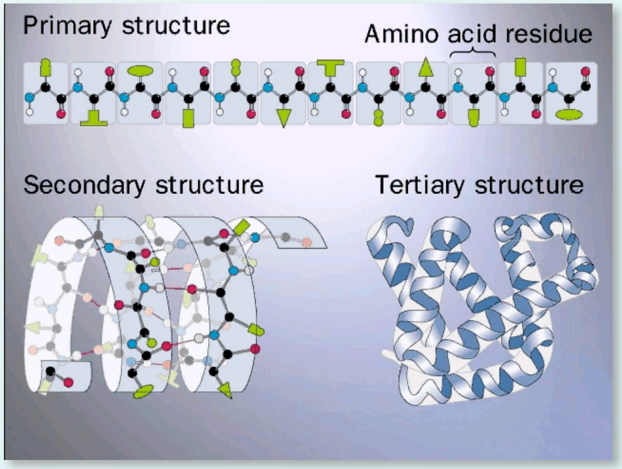
Protein Folding

Early in X-ray analysis, certain motifs appeared to occur frequently. The most common ones are the alpha-helix and the beta-sheet.





http://courses.bio.psu.edu/fall2005/biol230weve/tutorials/tutorial1.htm



This type of motif is referred to as a “secondary structure”.

The sequence along is the primary structure, and the almost folded protein is the “tertiary structure”, fully folded is “quaternary”.

Hydrophobic Effect

Most proteins usually have an interior composed of almost entirely nonpolar, hydrophobic amino acids.

This packing is a result of the hydrophobic effect, which is the most important factor that contributes to protein stability.

Water molecules exchange hydrogen bonds with neighbors at a rate of about 10^{11}/s.

At the interface between water and a non-H-bonding group such as CH_3, water molecules have fewer opportunities for H-bond exchange, leading to longer than usual lifetime of H-bonds, an ice-like state at the interface, and consequent decrease in entropy.

Any situation that minimizes the area of contact between H_2O and non-polar, i.e. hydrocarbon, regions of the protein results in an increase in entropy.

This is achieved by clustering nonpolar groups together.

First Principles Methods for Proteins

Solving the protein folding problem from first principles is the most ambitious approach.

The idea is to explore the conformational space of the molecule in order to identify the most appropriate structure.

Of course, the total number of structures is enormous, so usually one only tries to find the lowest energy structure.

By far the most common approach is to use an empirical force-field combined with some kind of solvation model (ie, no explicit water).

Once one finds a minimum-energy structure, one assumes it is the naturally occurring structure of that protein.

The problem is that conformational space is just too big to sample using any of the straightforward methods we've discussed so far…

Conformational Searches

Scheraga has been a big contributor to developing methods for exploring conformational space of proteins.

One of the popular ones is the “build-up” approach, where the protein is constructed from 3-dimensional amino acid templates.

Each template corresponds to a low-energy structure for a given sequence.

One firsts joins together all possible pairs of templates, with a minimization of the energy of the combined structure.

The lowest-energy joined structures are kept for the next step, which is connect a 3rd amino acid group.

The protein is gradually built up in this manner.

Conformational Searches

Another approach is based on Monte Carlo sampling.

Here, some structural property is varied, say the dihedral angles.

At each iteration, these angles are randomly rotated.

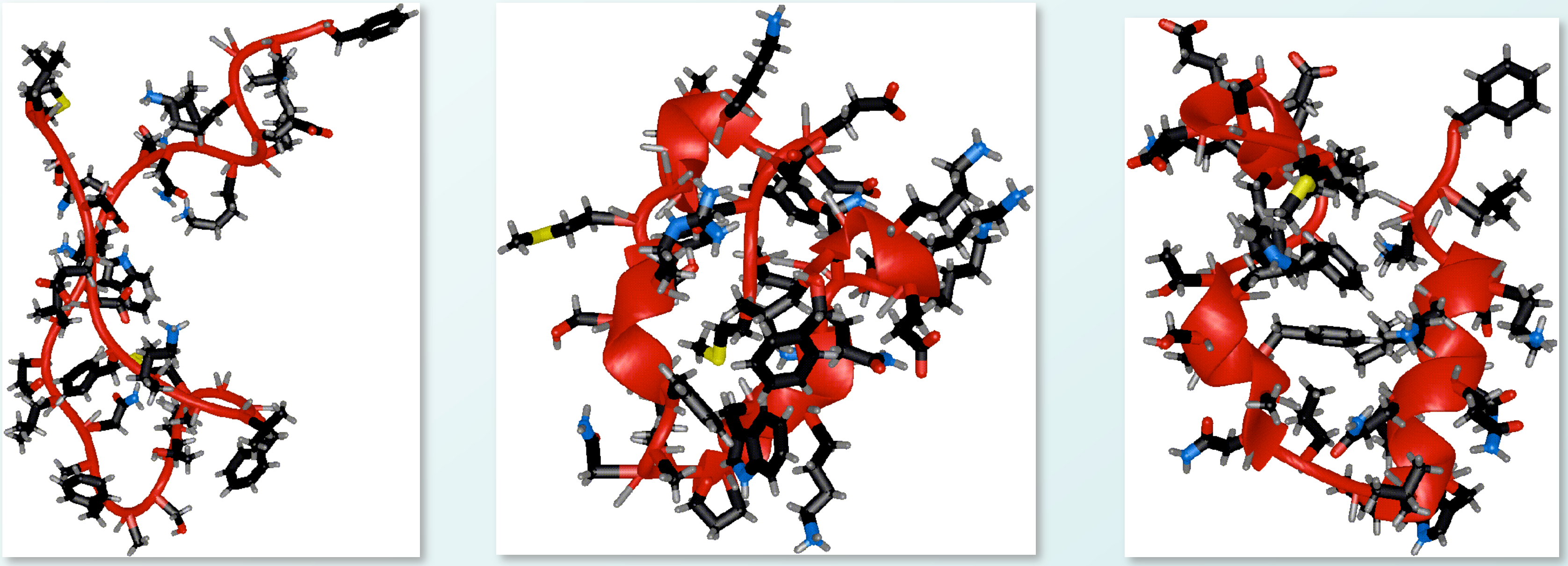
A Metropolis (you all know what this is right!?) criterion is used to accept or reject the move.

In some cases, such as the “electrostatically drive Monte Carlo approach” more complex interactions are accounted for (E field of protein interacting with dipoles of amino groups, for example).

Often the potential energy landscape is “smoothed” before any MC is carried out - this means that the number of degrees of freedom is reduced but the main features of the energy surface are reproduced.

Folding by Conformational Search

Here’s a random folding picture (cbcg.lbl.gov/ssi-csb/Chapter4.html)



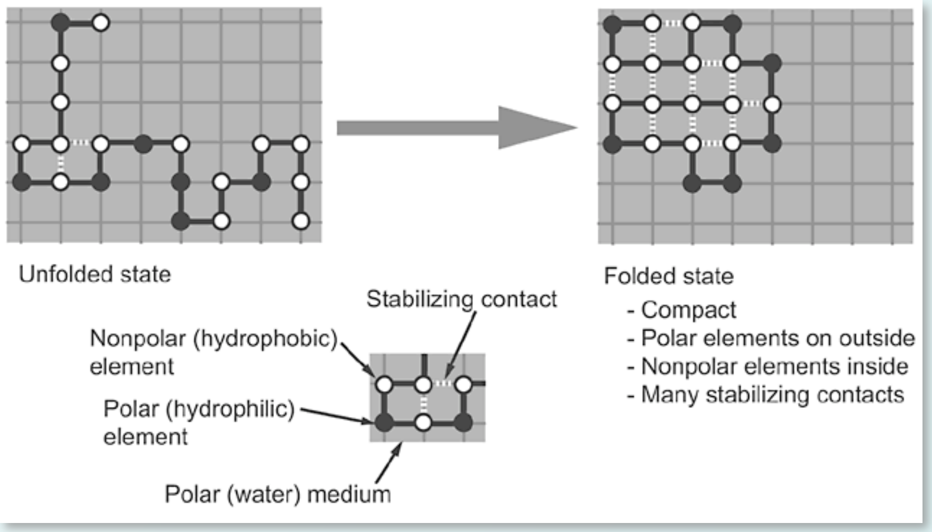
This took 2 1/2 months to simulate - a full microsecond.

Only the middle, metastable structure was obtained, as opposed to the final native state (right).

Lattice Models

Simple lattice models can capture important effects.

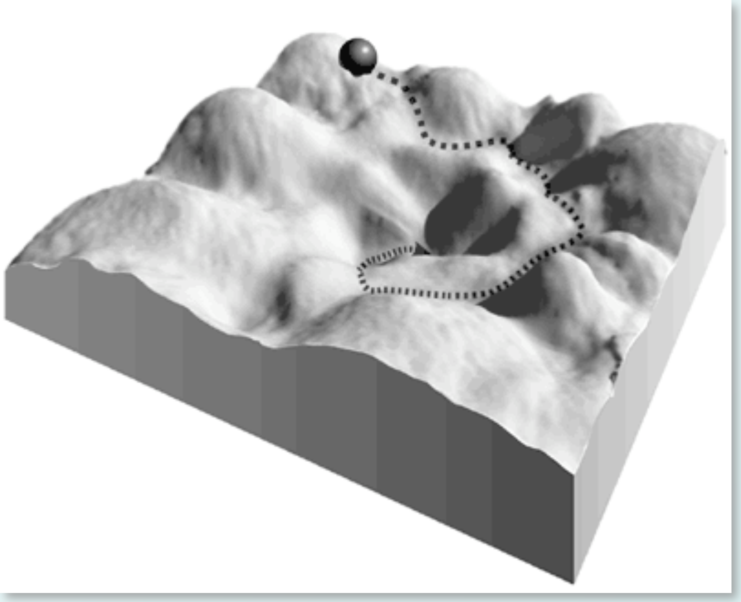
For example, if each amino acid is represented only by a bead on a lattice, and this bead is only hydrophobic or hydrophilic, then hydrophobic groups will favor positions wherein they are grouped together.


Unfolded state
Stabilizing contact
Nonpolar (hydrophobic) element
Polar (hydrophilic) element
Polar (water) medium
Folded state
- Compact
- Polar elements on outside
- Nonpolar elements inside
- Many stabilizing contacts

www.press.uillinois.edu/epub/books/brown/ch7.html

Levinthal Paradox

In 1969, Levinthal recognized that the protein folding process cannot be like a marble rolling through a hilly landscape as it samples energy configurations on its way to the lowest-energy state.



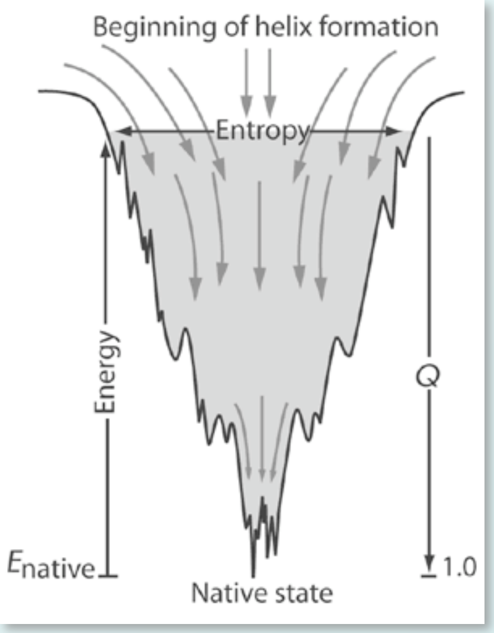
This would require the protein to sample too many configurations in a step-by-step process from unfolded to folded - it would take the age of the universe.

Levinthal Paradox

Instead, the landscape is considered to be much more “rugged” or “rough”.

The reason for this is that as the protein folds, certain parts of its sequence quickly find a “native” state, and when they do, they tend to stay in that stable state while the rest of the protein explores conformations.

The landscape is therefore much more like a funnel.



Rule-Based Approaches for Secondary Structures

The funnel energy landscape is a result of the fact that most protein structures contain a significant amount of secondary structure.

One approach, a rule-based approach, is to assign each amino acid in the sequence a secondary structure: alpha-helix, beta-strand, or neither (coil).

Secondary structure prediction can be done in many ways, but even the best approaches to date are only about 70% accurate (33% being random!).

The difficulty is that secondary structure prediction ignores interactions between amino acids that are far away in the sequence, but close in 3-D folding space.

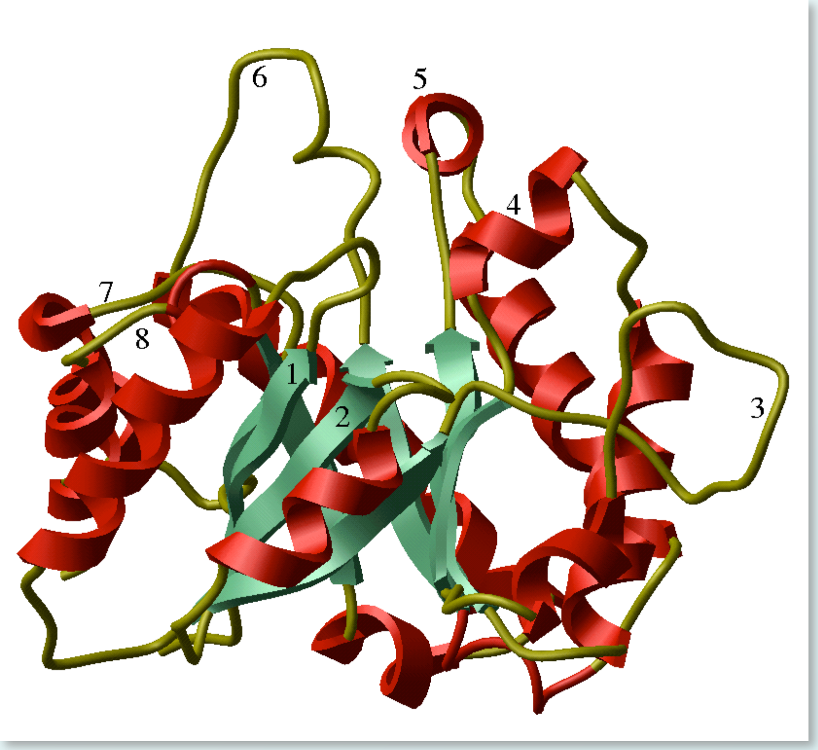
Once the secondary structures are predicted, it is then necessary to determine how they could pack together. A series of rules have been developed by, e.g., Cohen, et al.

These rules are based on simple, empirical observations of what usually happens in protein structures (e.g., alpha-helix packs against beta-sheet in parallel involving 2 rows of non-polar residues on the helix, etc.).

Comparative Modeling

Many proteins have very similar conformations.

For example, the “TIM barrel” contains 8 twisted parallel beta-strands arranged in a barrel-like structure with the beta-strands connected by alpha-helices.



www.biochem.oulu.fi/ tutkimus/wierenga/

Comparative Modeling

Comparative modeling exploits the structural similarities between proteins by constructing a 3-D structure based on known structure(s) of one or more related proteins.

Two important decisions are involved:

1. Which protein structure to use as the 3D template
2. How to match the amino acids in the unknown structure with those in the known structure

If the biological function of the protein is known, these decision can be straightforward.

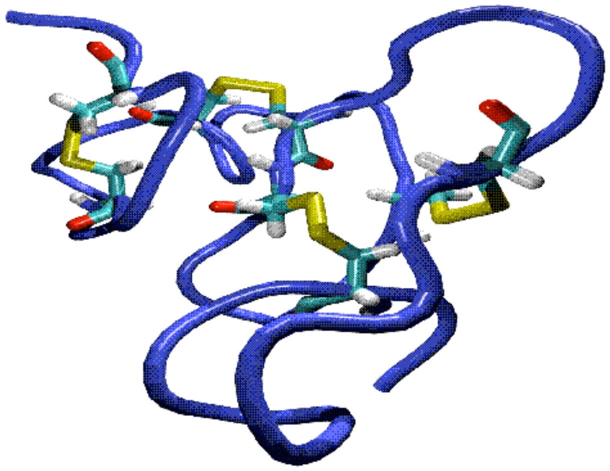
If the function is not known, one must search a sequence database for certain combinations of amino acids that imply a particular function or structure.

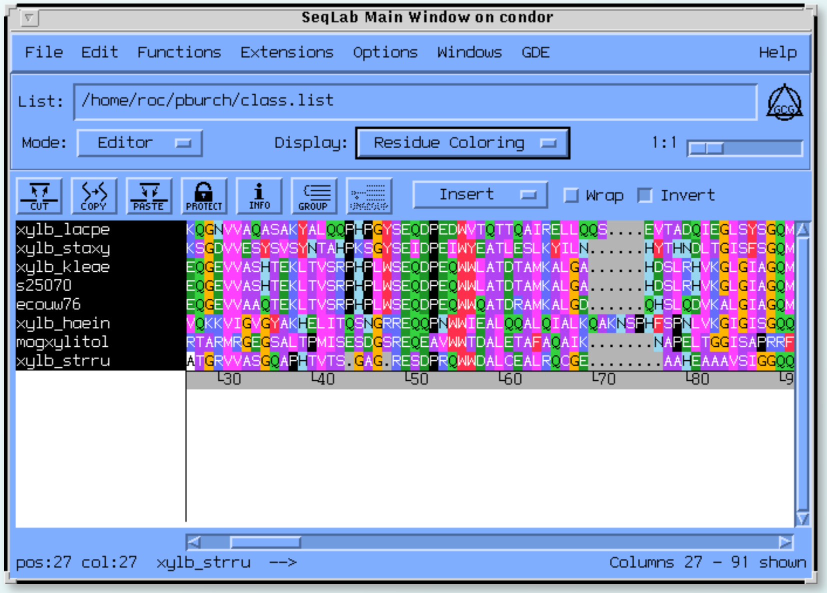
Sequence Alignment

Very simply, proteins with similar sequences tend to have similar structures.

The objective of a sequence alignment algorithm is to position the amino acid sequences so that the matched sections correspond to common structural or functional features (e.g., secondary structures).

```
AATAHAQRCG EQGSNMECPN NLCCSQYGYC GMGGDYCGKG ..CQNGACYT
VAATNAQTCG KQNDGMICPH NLCCSQFGYC GLGRDYCGTG ..CQSGACCS
VGLVSAQRCG SQGGGGTCPA LWCCSIWGWC GDSEPYCGRT ..CENK.CWS
AATAQAQRCG EQGSNMECPN NLCCSQYGYC GMGGDYCGKG ..CQNGACWT
AATAQAQRCG EQGSNMECPN NLCCSQYGYC GMGGDYCGKG ..CQNGACWT
......QRCG EQGSGMECPN NLCCSQYGYC GMGGDYCGKG ..CQNGACWT
SETVKSQNCG .......CAP NLCCSQFGYC GSTDAYCGTG ..CRSGPCRS
RGSAE..QCG RQAGDALCPG GLCCSSYGWC GTTVDYCGIG ..CQSQ.CDG
AGPAAAQNCG .......CQP NFCCSKFGYC GTTDAYCGDG ..CQSGPCRS
AGPAAAQNCG .......CQP NVCCSKFGYC GTTDEYCGDG ..CQSGPCRS
RGSAE..QCG QQAGDALCPG GLCCSSYGWC GTTADYCGDG ..CQSQ.CDG
RGSAE..QCG RQAGDALCPG GLCCSFYGWC GTTVDYCGDG ..CQSQ.CDG
TGVAIAEQCG RQAGGKLCPN NLCCSQWGWC GSTDEYCSPD HNCQSN.CK.
......EQCG RQAGGKLCPN NLCCSQYGWC GSSDDYCSPS KNCQSN.CK.
```





List: /home/roc/pburch/class.list

Mode: Editor   Display: Residue Coloring   1:1

□ Wrap □ Invert

```
xylb_lacpe    KQGNVVAQASAKYALQQPHPGYSEQDPEDWVTQTTQAIRELLQQS.....EVTADQIEGLSYSGQM
xylb_staxy    KSGDVVESYSVSYNTAHPKSGYSEIDPEIWYEATLESLKYILN.......HYTHNDLTGISFSGQM
xylb_kleae    EQGEVVASHTEKLTVSRPHPLWSEQDPEQWWLATDTAMKALGA.......HDSLRHVKGLGIAGQM
s25070        EQGEVVASHTEKLTVSRPHPLWSEQDPEQWWLATDTAMKALGA.......HDSLRHVKGLGIAGQM
ecouw76       EQGEVVAAQTEKLTVSRPHPLWSEQDPEQWWQATDRAMKALGD.......QHSLQDVKALGIAGQM
xylb_haein    VQKKVIGVGYAKHELITQSNGRREQQPNWWIEALQQALQIALKQAKNSPHFSPNLVKGIGISGQQ
mogxylitol    RTARMRGEGSALTPMISESDGSREQEAVWWTDALETAFAQAIK........NAPELTGGISAPRRF
xylb_strru    ATGRVVASGQAPHTVTS.GAG.RESDPRQWWDALCEALRQCGE.........AAHEAAAVSIGGQQ
                └30       └40       └50       └60       └70       └80       └9
```

pos:27 col:27   xylb_strru -->   Columns 27 - 91 shown

Gaps in the aligned sequences correspond to regions where polypeptide loops are inserted or deleted.

Sequence Alignment

There are a number of general sequence alignment schemes.

Attempts to match the full length of the sequence (e.g., Dayhoff et al), give alignment “scores” but take into account complex interactions within the sequence (even including mutations).

Approaches based on local alignments (e.g., Needleman and Wunsch) are based on complex “dynamic programming” algorithms which are based on the construction and evolution of a scoring matrix.

The most promising newer approaches are heuristic and much more suited to search the ever-increasing database of sequences. (see, “FASTA” and “BLAST” methods).

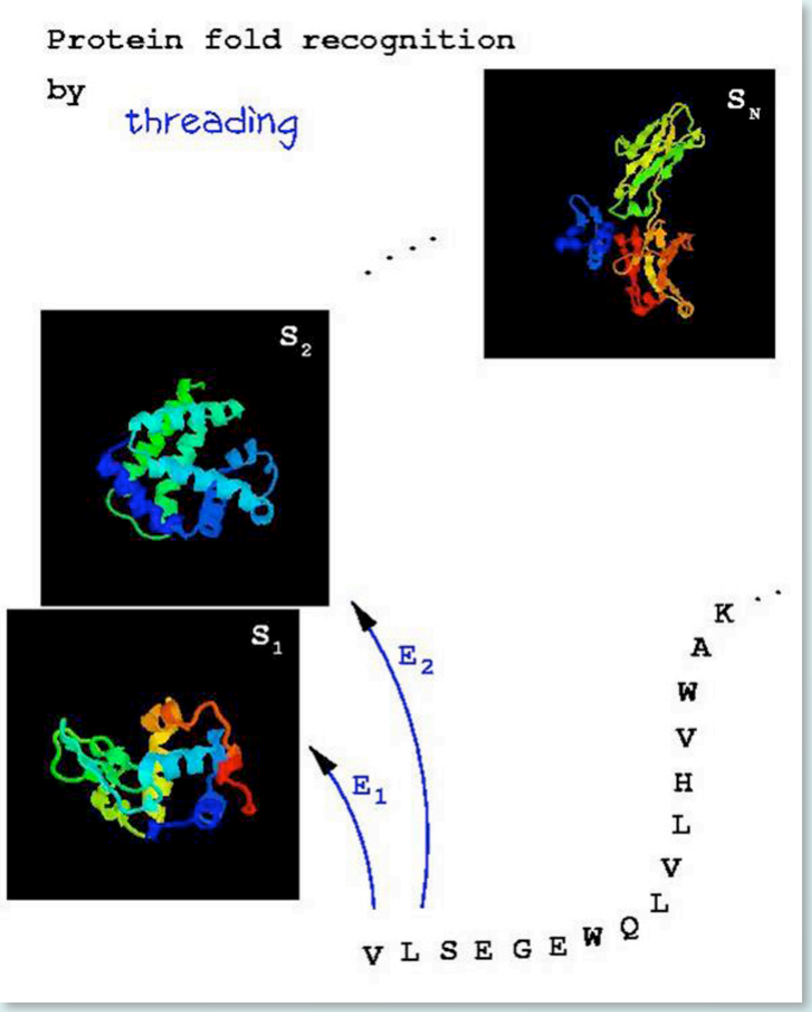
Multiple sequences, which involve comparing with more than 1 other sequence are more reliable.

Threading

If the sequence identity between two proteins is below 30% then comparative modeling is not possible.

One alternative is called “threading”. Here several conformers are used to thread the amino acid sequence through.

At each thread, a score is given and the amino acid sequence is moved up one position and a new score is made.



www.cincinnatichildrens.org/.../meller/tools/