

Coarse grain models (continuous)

Andrew Torda, june 2008, 67.912, Strukt & Sim

What have we seen so far ?

- very detailed models
 - QM
 - atomistic, solvation

What are some reasonable aims ?

- given a set of coordinates
 - are these roughly correct for a protein sequence ?
 - is this more likely to be α -helical or β -sheet ?
- less reasonable
 - given initial coordinates, can I simulate a protein folding ?

Should we approach this with a detailed force field ?

- maybe not

Aims

- Why atomistic force fields / score functions are not always best
- Different levels of force fields
- Examples of coarse-grain / low-resolution force fields
- Ways to parameterise force fields

- later...
- extending this idea to lattice models

History

History

- Levitt, M and Warshel, A, Nature, 253, 694-698, Computer simulation of protein folding (1975)
- Kuntz, ID, Crippen, GM, Kollman, PA and Kimelman, D, J. Mol. Biol, 106, 983-994, Calculation of protein tertiary structure (1976)
- Levitt, M, J. Mol. Biol, 104, 59-107, A simplified representation of protein conformations for rapid simulation of protein folding (1976)
- through to today

Problems with detailed force fields

Time

- typical atomistic protein simulations 10^{-9} to 10^{-6} s
- too short for folding

Radius of convergence

- I have coordinates where atoms are perturbed by 1 Å
 - easy to fix – atoms move quickly
- I have completely misfolded, but well packed coordinates
 - may be difficult to fix
 - what dominates ?
 - atomic packing
 - charges
 - solvation ?

Do I care about details ?

Coarse grain / low resolution

Forget atomic details

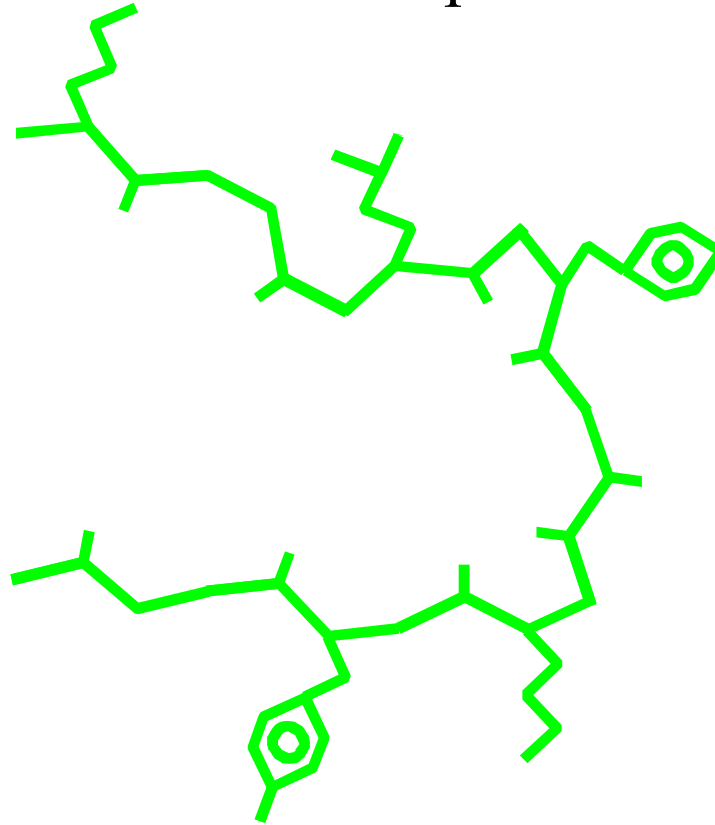
- build something like energy which encapsulates our ideas
- example – define a function which is happiest with
 - hydrophobic residues together
 - charged residues on outside
- would this be enough ?
 - maybe / not for everything

What will I need ?

- some residues like to be near each other (hydrophobic)
- residues are always some constant distance from each other
- only certain backbone angles are allowed

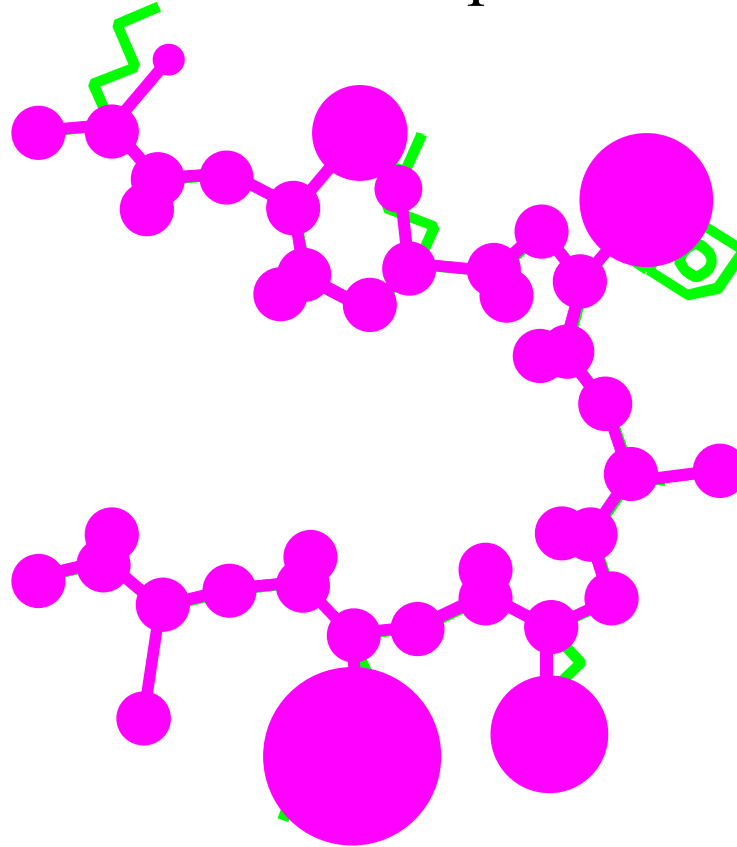
General implementation (easiest)

- how do we represent a protein ?
 - decide on number of sites per residue



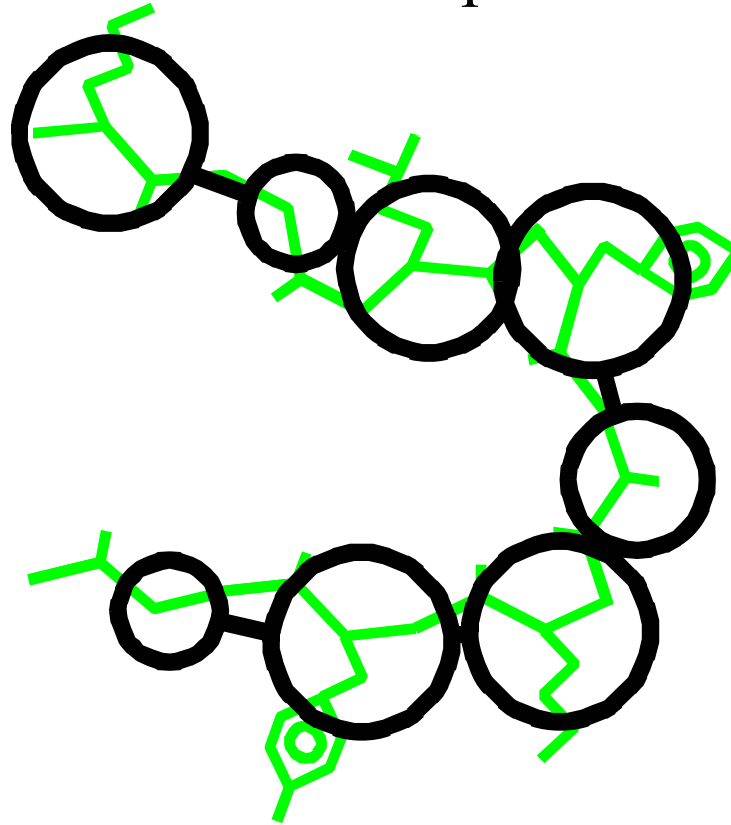
General implementation (easiest)

- how do we represent a protein ?
 - decide on number of sites per residue



General implementation (easiest)

- how do we represent a protein ?
 - decide on number of sites per residue



Coarse-graining (steps)

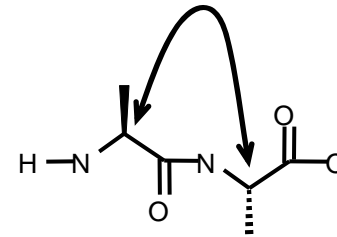
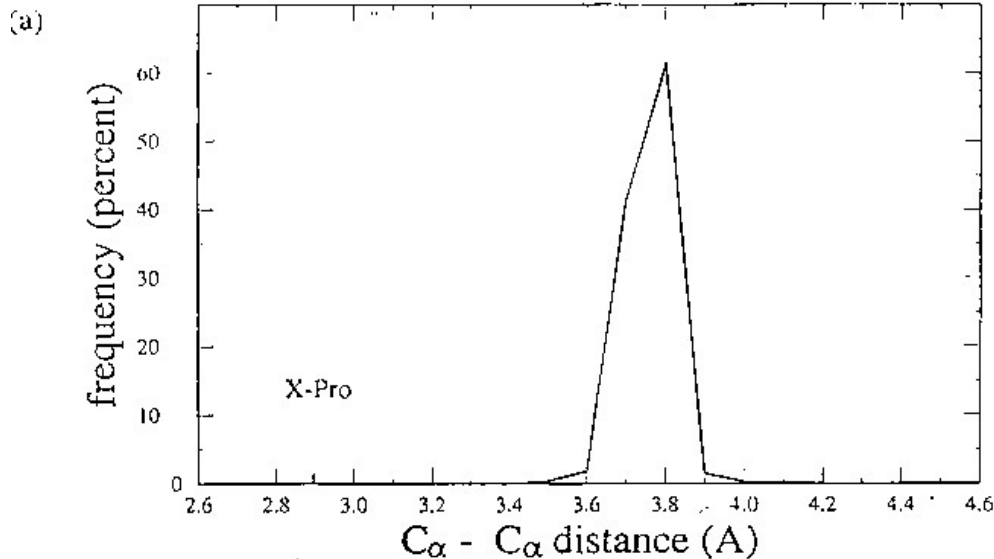
- Decide on representation
- Invent quasi-energy functions
- Our plan
 - step through some examples from literature

Common features

- some way to maintain basic geometry
- size
- hydrophobicity ? which residues interact with each other/solvent

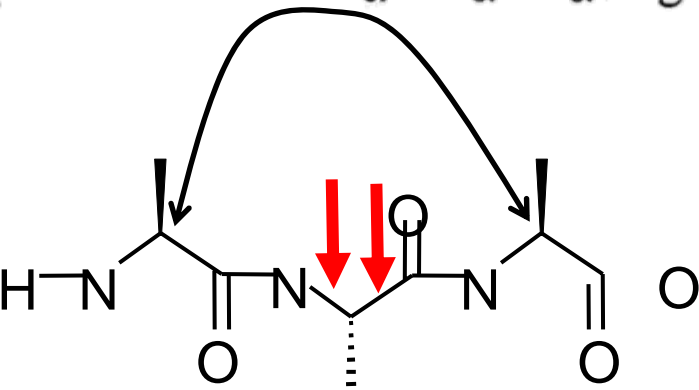
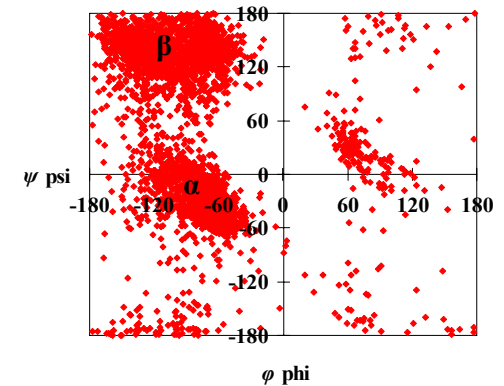
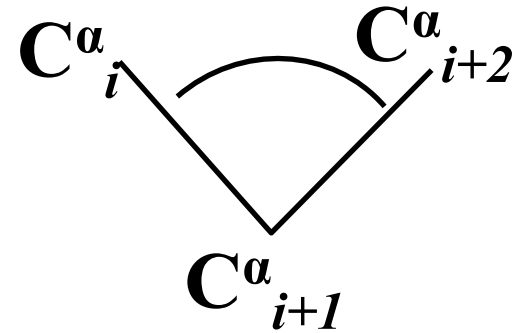
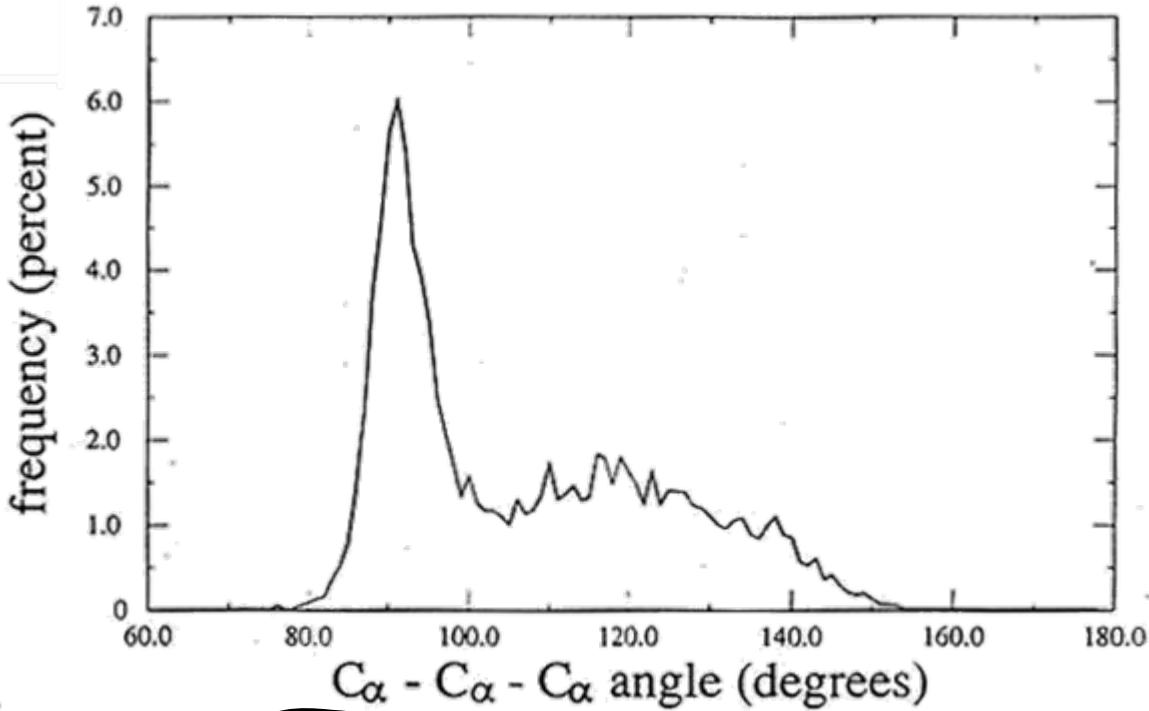
Basic geometry

- Survey protein data bank files and look at C^α to C^α distances



- Conclusion is easy
 - any model should fix $C^\alpha_{i,i+1}$ distances at 3.8 Å
- what other properties do we know ?

$C^\alpha_{i,i+2}$ distance / angle



- why is distance less clear ?
- think of ramachandran plot

First simple model

n residues, n interaction sites $i, i+1$ restrained (C^β formulation)

Overlap penalty / radii

- lys 4.3 Å, gly 2.0 Å, ... trp 5.0 Å
- $U(r_{ij}) = (\text{radius}_i + \text{radius}_j)^2 - r_{ij}^2$

force hydrophilic residues to surface, for these residues

- $U^*(r_{ij}) = (100 - d_i^2)$ where d_i is distance to centre, 100 is arbitrary

disulfide bonds

- very strong

residue specific interactions

- $U^{long}(r_{ij}) = c_{ij} (r_{ij}^2 - R^2)$ where c_{ij} is residue specific
- R is 10 Å for attraction, 15 Å for repulsion

residue specific part of interaction

- c_{ij} table
- features
 - hydrophobic
 - + -
 - nothing much

	lys	glu	...	gly	pro	val
lys	25	-10		0	0	10
glu	-10	25		0	0	10
...						
gly	0	0		0	0	0
pro	0	0		0	0	0
val	10	10		0	0	-8

summary

- $i, i+1$ residue-residue
- overlap
- long range
- solvation

where is physics ?

- solvation ?
 - term pushes some residues away from centre
- electrostatics
- hydrophobic attraction
 - by pair specific c_{ij} terms

other properties

- smooth / continuous function
- derivative with respect to coordinates
 - (good for minimisation)

does it work ? what can one do ?

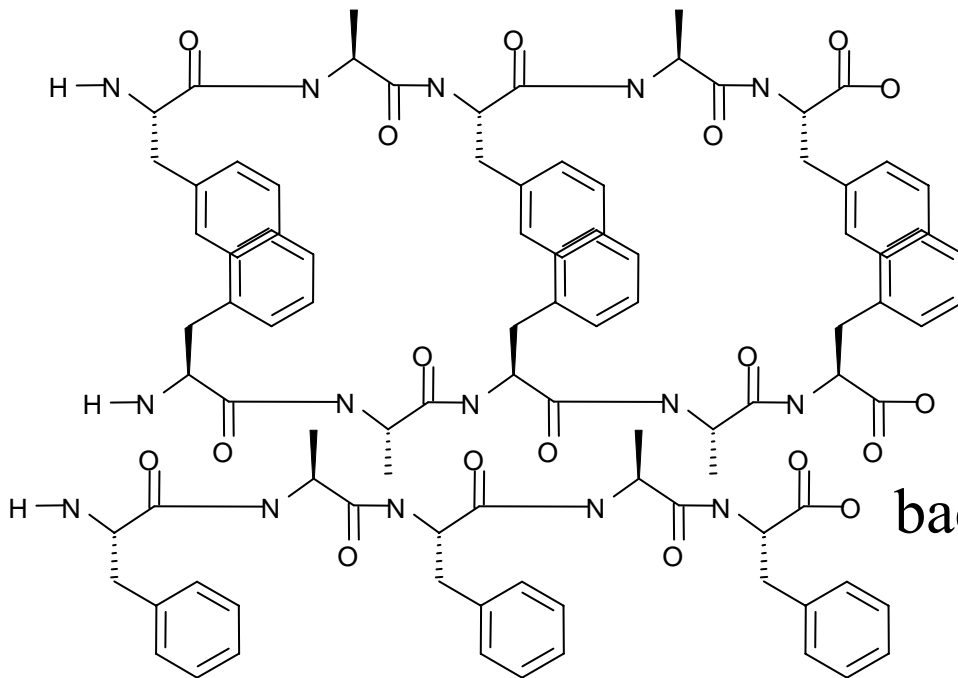
results from first model

- try to "optimise" protein structure
- for 50 residues, maybe about 5 Å rms
 - maybe not important
- model does..
 - make a hydrophobic core
 - put charged and polar residues at surface
 - differentiate between possible and impossible structures
- model does not
 - reproduce any geometry to Å accuracy
 - details of secondary structure types
 - not the intention
 - predict physical pathways
 - depend on subtle sequence features (simplicity of c_{ij} matrix)

Improvements to simple model

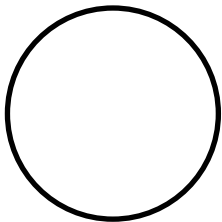
- aim
 - biggest improvement for least complication
- possibilities
 - more points per residue
 - more complicated c_{ij} matrix...
 - an example weakness
- important structural features of proteins
 - all proteins have hydrogen bonds at backbone
 - proteins differ in their sidechain interactions..

more complicated interactions

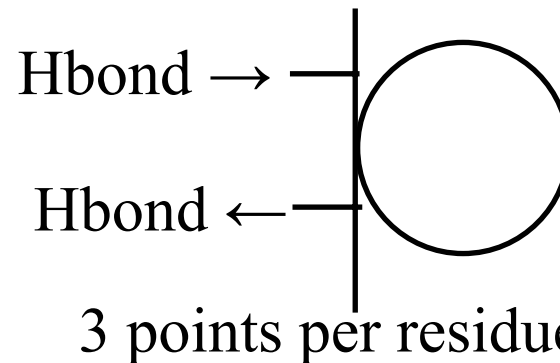


sidechain packing

backbone Hbonds



one point residue



Hbond →

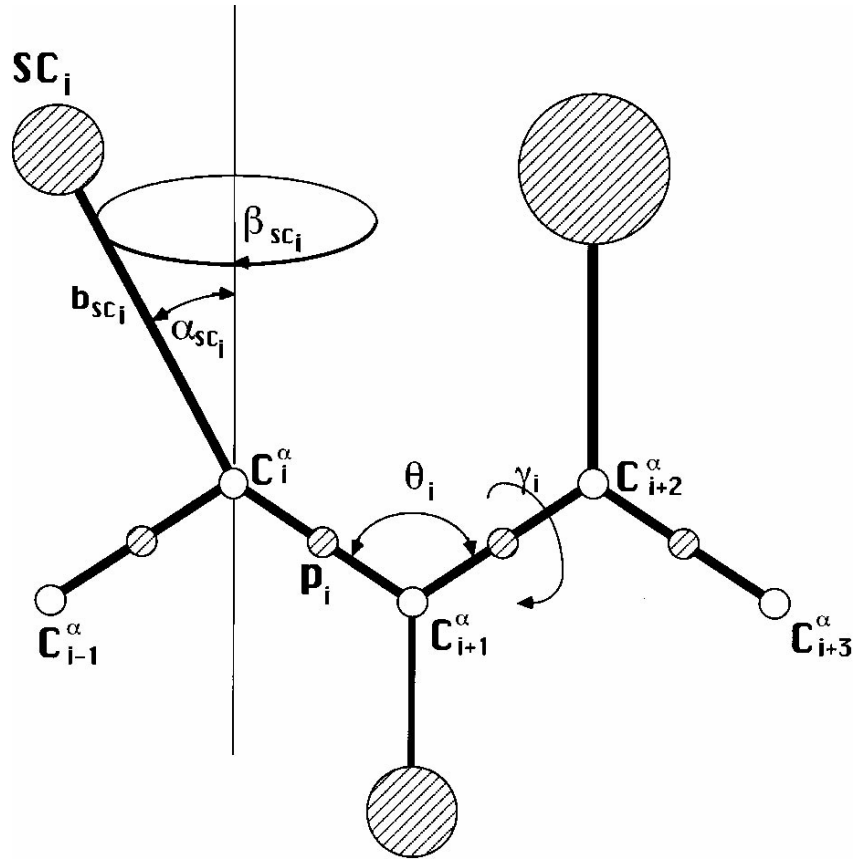
Hbond ←

3 points per residue

Scheraga model

3 points per residue

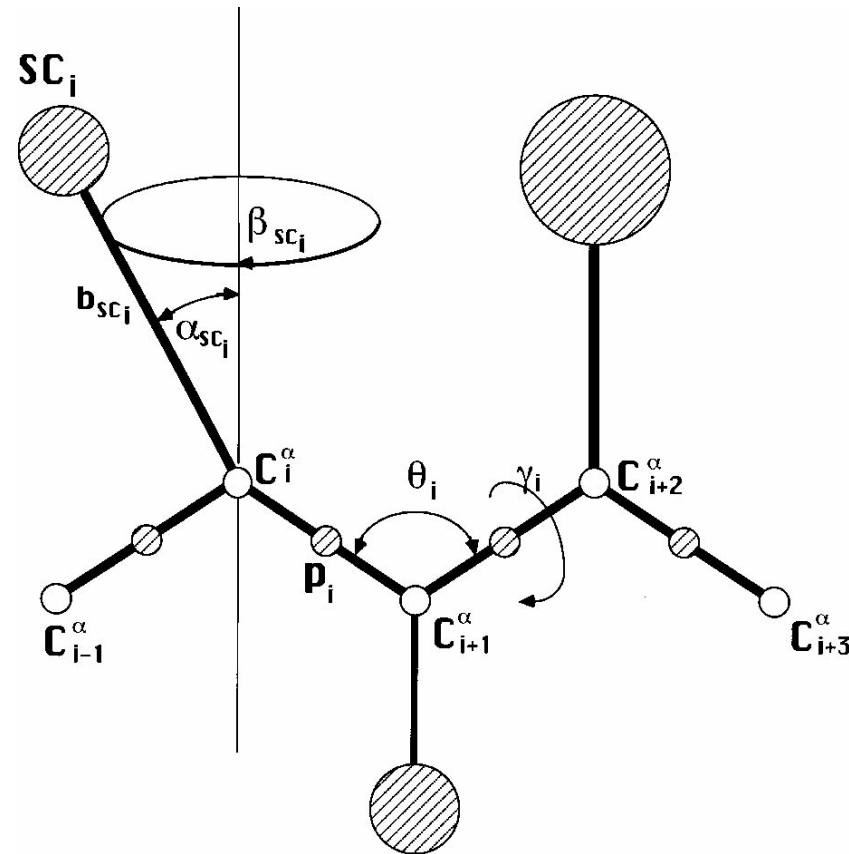
- 2 for interactions
 - p_i is peptide bond centre
 - SC_i is sidechain
- 1 for geometry
 - C^α
- $C^\alpha - C^\alpha$ fixed at 3.8 \AA



- do interaction sites correspond to atoms ?

Terms in Scheraga model

- Total quasi energy =
 - side-chain to side-chain
 - side-chain to peptide
 - peptide to peptide
 - torsion angle γ
 - bending of θ
 - ...
 - bending α_{sc}

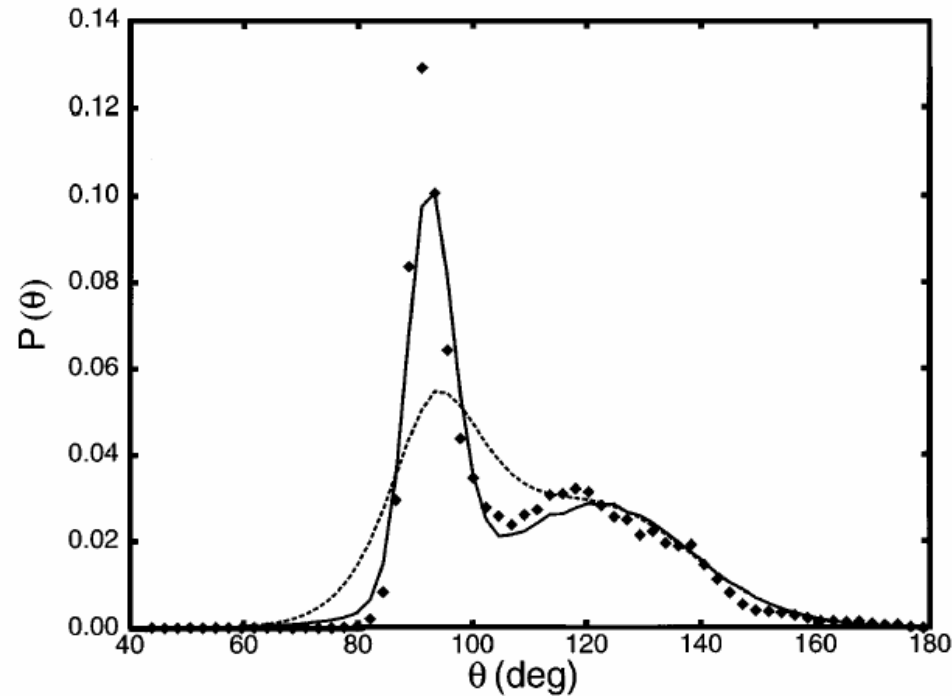
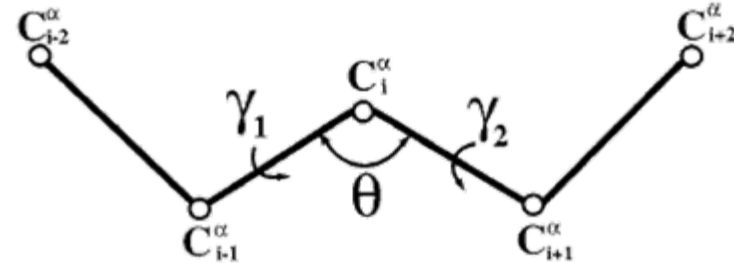


angle between C^α sites

- cunning approach
 - look at θ distribution
 - model with Gaussians + frills (solid line)
- then say

$$U(\theta)^{bend} = -RT \log P(\theta)$$

- where $P(x)$ is the probability of finding a certain x



Gaussian reminder

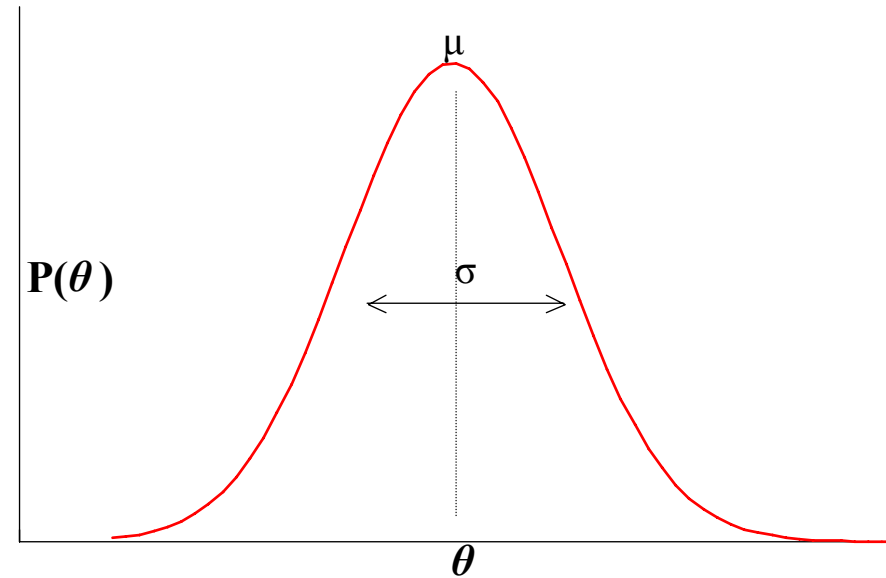
- get μ and σ from fitting
- angle θ depends on structure

$$P(\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(\theta - \mu)^2}{2\sigma^2}\right]$$

- how would forces work ?
- express θ in terms of r 's
- use $U(\theta)^{bend} = -RT \log P(\theta)$

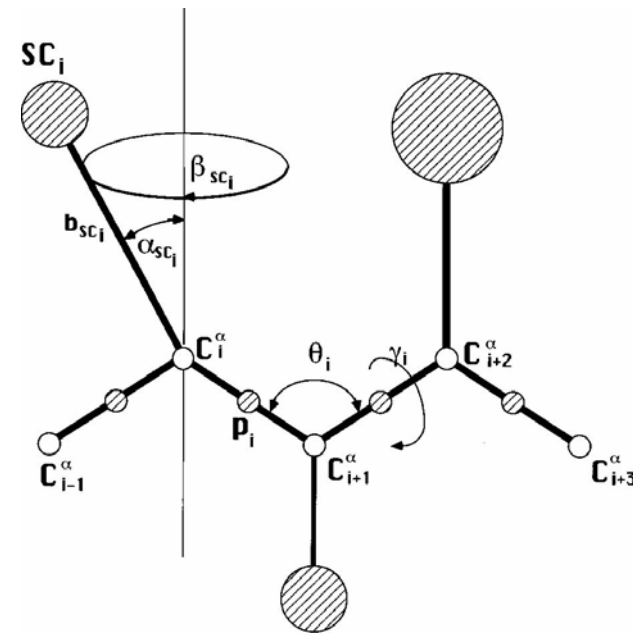
- take $\frac{dU}{d\theta} \frac{\partial \theta}{\partial \vec{r}}$

- first part ugly, second part ugly



pseudo torsion term

- like an atomic torsion $U(\gamma_i) = a_i(\cos n\gamma_i + 1) + b_i(\sin n\gamma_i + 1)$
 - n varies from 3 to 6 depending on types i, j
 - three kinds of i, j pair
 - gly
 - pro
 - others
- net result ?
 - residues will be positioned so as to populate correct parts of ramachandran plot
 - this model will reproduce α -helix and β -sheets



side-chain peptide

- maybe not so important
 - mostly repulsive $U^{sc-peptide}(r_{SCp}) = kr_{SCp}^{-6}$
 - k is positive, so energy goes up as particles approach

side chain interactions

Familiar $U(r_{ij}) = 4\varepsilon_{ij}(\sigma_{ij}r_{ij}^{-12} - \sigma_{ij}r_{ij}^{-6})$

- but, consider all the σ and ε
- main result
 - some side chains like each other (big ε)
 - some pairs can be entirely repulsive (small ε big σ)
 - some not important (small ε small σ)

more complications

- real work used
 - different forms for long range interactions
 - cross terms in pseudo angles

What can one do ?

Typical application

Background

- protein comparison lectures..
- different sequences have similar structure
 - can we test some structure for a sequence

Remember sequence + structure testing in Übung ?

- here
 - given some possible structures for a sequence
 - can be tested with this simple force field

What can we not do ?

- physical simulations
 - think of energy barriers (not real)
 - time scale

summary of philosophy

- Is any model better than others ?
- Each model has represent something of interest
 - hydrophobic / hydrophilic separation
 - reasonably good quality structure with
 - real secondary structure
 - accurate geometry
- Main aims
 - pick the simplest model which reproduces quantity of interest
- Are there bad models ?
 - complicated, but not effective
 - interaction sites at wrong places
 - not efficient
 - not effective

Parameterisation..

Problem example

- charge of an atom ?
 - can be guessed, measured ?, calculated from QM
- ϵ and σ in atomistic systems
 - can be taken from experiment (maybe)
 - adjust to reproduce something like density

What if a particle is a whole amino acid or sidechain ?

- is there such a thing as
- charge ?
- ϵ and σ ?

Approaches to parameterisation

General methods

- average over more detailed force field (brief)
- optimise / adjust for properties (brief)
- potentials of mean force / knowledge based (detailed)

From detailed to coarse grain

Assume detailed model is best

- Can we derive coarse grain properties from detailed ?

Examples – consider one or two sites per residue

- mass ? easy – add up the mass of atoms (also boring)

- charge ? not easy

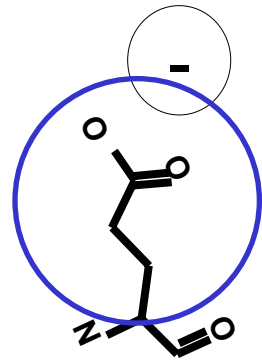
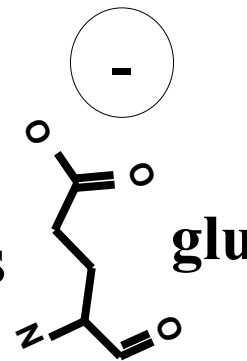
- is charge important ? sometimes

- size of charge obvious

- location of charge may not be the same as a single site

- does this let us include polarity ? No.

- is this the right way to think about it ?...



Averaging over details is not easy

If we have electrostatics

- perhaps we can have coarse electrostatics
- maybe better to forget serious physics / strict electrostatics

Earlier example (Kuntz et al)

- pairwise interactions like $(r_0^2 - r_{ij}^2)$ + term for sending residues / to from centre of molecule
- you can not easily get parameters from a more detailed force field here

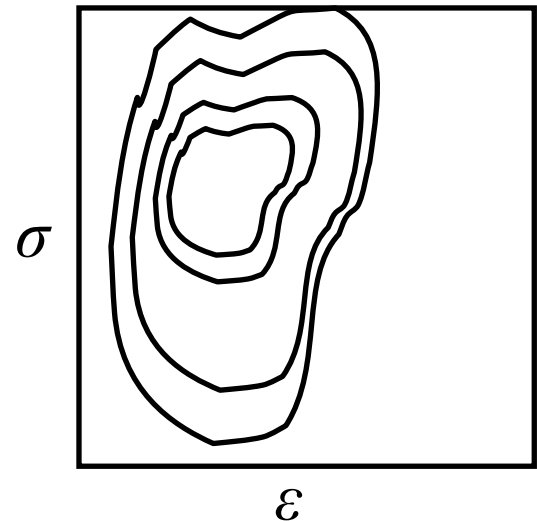
General interaction between two residues

- will depend on orientation, distance, other neighbours
- not all orientations occur equally likely
- sensible averaging not obvious
- better approach ...

Parameterising by adjustment

Basic idea

- build some representation (like examples above)
- adjust parameters to give desired result
- An example method
- define a simple force field like $U(r_{ij}) = 4\varepsilon_{ij} (\sigma_{ij} r_{ij}^{-12} - \sigma_{ij} r_{ij}^{-6})$
 - run a calculation and measure a property
 - density ? how near to correct structure ?
 - repeat for many values of ε and σ
 - build a cost / merit map



mapping parameter space

What does this tell us ?

- pinpoint the best ε and σ
- see that ε is critical, σ less so

Good result ?

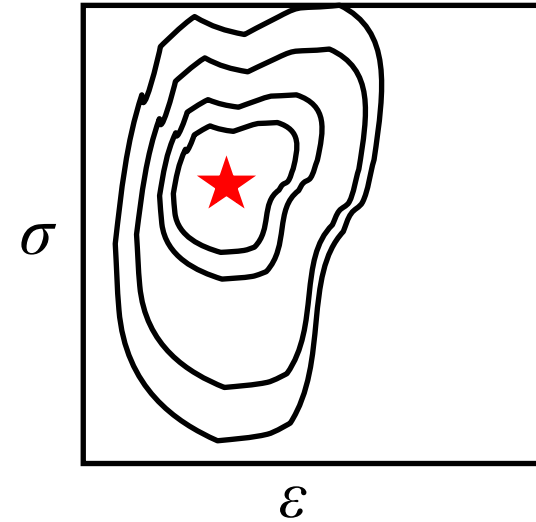
- parameters from one or several proteins should work on all

Refinement ?

- optimisation can be automated

Problems

- scheme requires a believable measure of quality
- easy for two parameters
- possible for 3, 4 parameters
- very difficult for 100 parameters



parameterising from potential of mean force

Potential of mean force ... knowledge based score functions

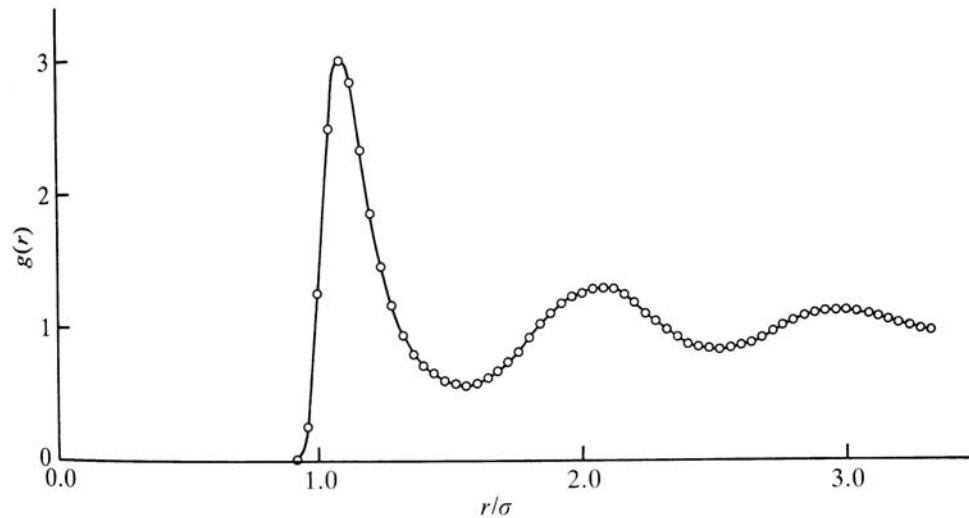
- very general
- history from atomistic simulations

Basic idea .. easy

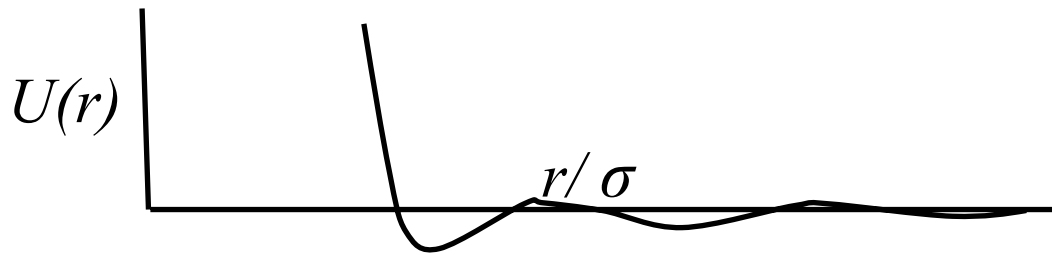
- from radial distribution function, to something like energy..

Intuitive version of potential of mean force

- radial distribution function $g(r)$
 - probability of finding a neighbour at a certain distance



- what does this suggest about energy ?



Radial distribution function

- Formal idea

$$g(r) = \frac{N_{\text{neighbours seen}(r)}}{N_{\text{neighbours expected}(r)}}$$

- N particles

- V volume

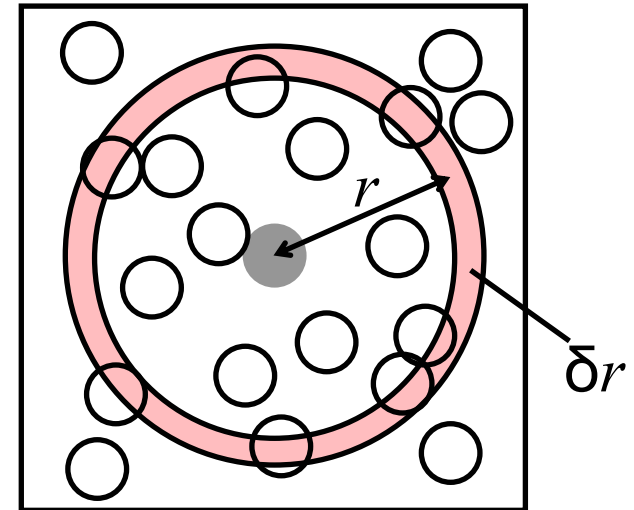
$$N_{\text{expected}} = \frac{V_{\text{shell}}}{V} N$$

- Calculating it ?

- define a shell thickness (δr)

- around each particle

- at each distance, count neighbours within shell



$$g(r) = \frac{V}{NV_{\text{shell}}} N_{\text{shell}}(r)$$

Rationale for potentials of mean force

- For state i compared to some reference x

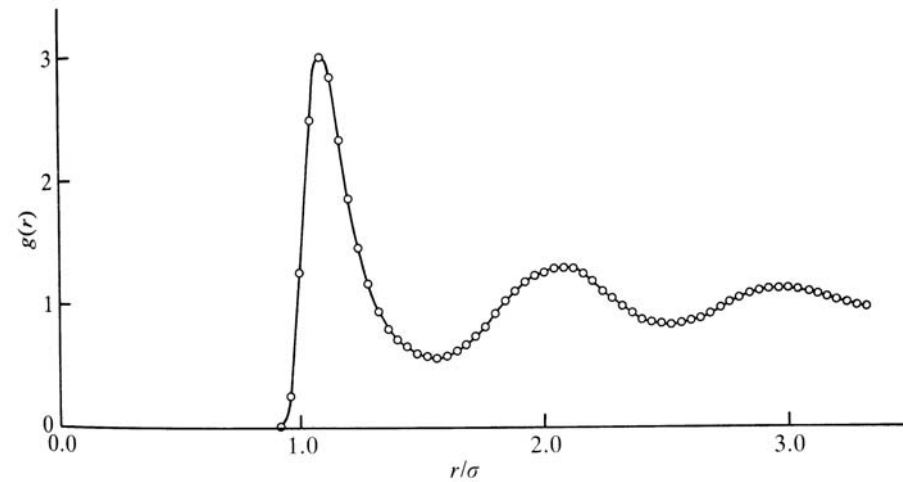
$$\frac{p_i}{p_x} = \frac{e^{-E_i/kT}}{e^{-E_x/kT}}$$

$$= e^{E_x - E_i / kT}$$

$$\ln \frac{p_i}{p_x} = \frac{E_x - E_i}{kT}$$

$$\Delta E = kT \ln \frac{p_i}{p_x}$$

Information in distribution function



Intuitive properties ?

- how likely is it that atoms get near to each other ($< \sigma$) ?
- what would a crystal look like ? (very ordered)
- what if interactions are
 - very strong (compared to temperature)
 - very weak
- Seems to reflect all properties of a system
 - strength of interactions / order

Relate this back to energy

Energy from $g(r)$

from statistical mechanics $g(r) = e^{-w(r)/kT}$

- use work $w(r)$ for a picture moving particle by r

so strictly $w(r) = -kT \ln g(r)$

- already useful for looking at liquid systems
- properties
 - are we looking at potential energy U or free energy G ?
 - if our results from nature (or simulation) – free energy
- how would we get $g(r)$?
 - experiment ? sometimes
 - simulation – easy
- assumptions
 - our system is at equilibrium
 - it is some kind of ensemble

Generalising ideas of potential of mean force

What else can we do ?

- think of more interesting system (H_2O)

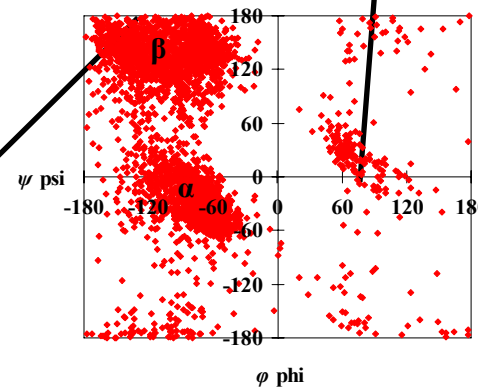
Would we express our function in terms of O ? H ?

- both valid
- could look at the work done to bring an O to O, O to H, H to H

More general..

- are we limited to distances ? No
- example – ramachandran plot

high probability /
low energy



low probability /
high energy

reformulating for our purposes

Can one use these ideas for proteins ?

Our goal ?

- a force field / score function for deciding if a protein is happy
- work with particles / interaction sites
- slightly different formulation
 - if I see a pair of particles close to each other,
 - is this more or less likely than random chance ?
 - treat pieces of protein like a gas
 - care about types of particles (unlike simple liquid)
- Let us define...

Score energy formulation

Define

$$W_{AB}(r) = -RT \ln \left(\frac{N_{AB}^{obs}(r \pm \delta r)}{N_{AB}^{exp}(r \pm \delta r)} \right)$$

- N_{AB}^{obs} how many times do we see
 - particles of types A and B
 - distance r given some range δr
- N_{AB}^{exp} how often would you expect to see AB pair at r ?
- remember Boltzmann statistics

This is not yet an energy / score function !

- it is how to build one

Intuitive version

- Cl^- and Na^+ in water like to interact (distance r^0)
- N_{AB}^{obs} is higher than random particles
- $W_{\text{ClNa}}(r)$ is more negative at r^0

Details of formulation

$$W_{AB}(r) = -RT \ln \left(\frac{N_{AB}^{obs}(r \pm \delta r)}{N_{AB}^{exp}(r \pm \delta r)} \right)$$

- looks easy, but what is N^{exp} ?
- maybe fraction of particles is a good approximation
 - $N_{NaCl}^{exp} = N_{alt} X_{Na} X_{Cl}$ (use mole fractions)
- use this idea to build a protein force field / score function

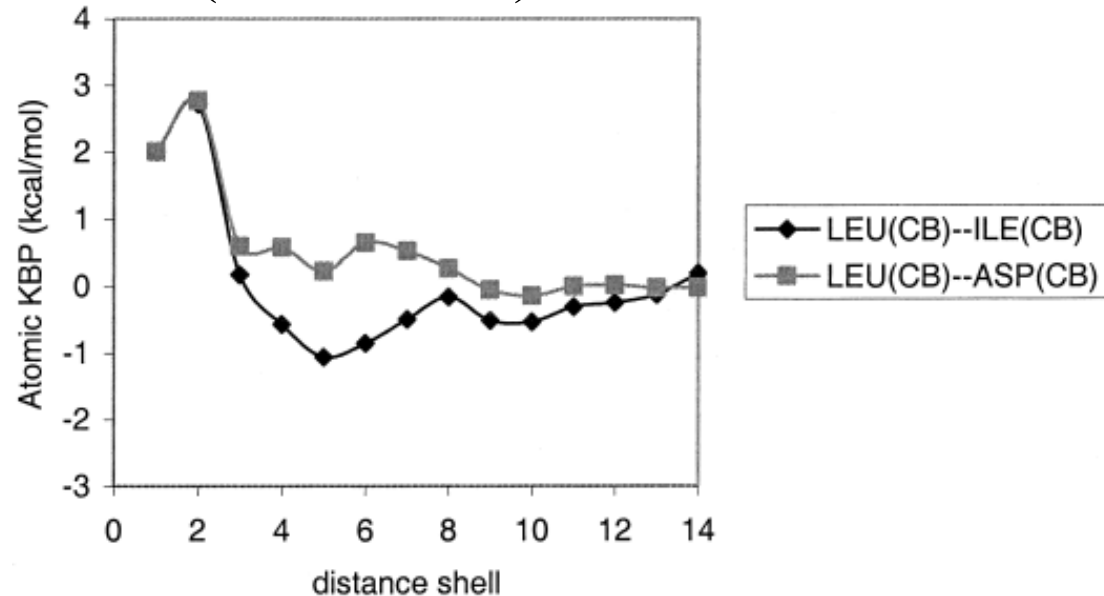
Protein score function

Arbitrarily

- define interaction sites as one per residue
 - maybe at C^α or C^β
- collect set of structures from protein data bank
- define a distance (4 Å) and range (± 0.5 Å)
- count how often do I see
 - gly-gly at this range, gly-ala, gly-X, X-Y ...
 - gives me N^{obs}
 - how many pairs of type gly-gly, gly-ala, gly-X, X-Y... are there ?
 - gives me N^{exp}
 - repeat for 5 Å, 6 Å, ...
- resulting score function...

final score function

- for every type of interaction AB (20 x 21 / 2)
 - set of $W_{AB}(r)$



All ingredients in place

- can we use this for simulations ? not easy
- can we use to score a protein ? yes

Names

- Boltzmann-based, knowledge based

Applying knowledge-based score function

Take your protein

- for every pair of residues
 - calculate C^β C^β distance (for example)
 - look up type of residues (ala-ala, trp-ala, ...)
 - look up distance range
 - add in value from table
- what is intuitive result from a
 - a sensible protein / a misfolded protein ?
- is this a real force field ? yes
- is this like the atomistic ones ? no
 - there are no derivatives (dU / dr)
 - it is not necessarily defined for all coordinates

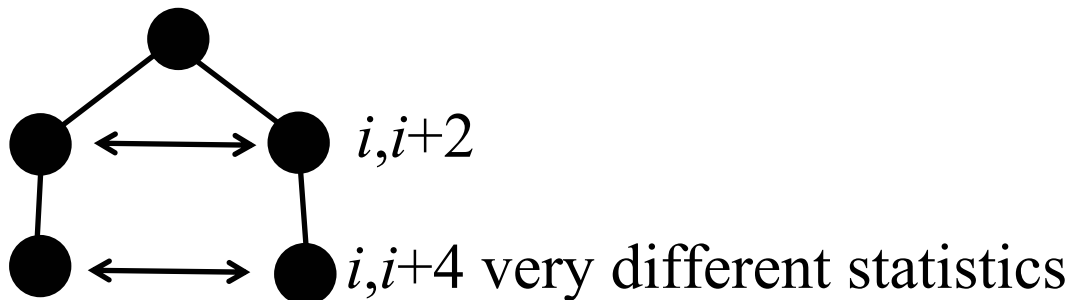
Practical Problems Boltzmann score functions

Practical

- Do we have enough data ?
 - how common are Asp-Asp pairs at short distance ?
- How should we pick distance ranges ?
 - small bins (δr) give a lot of detail, but there is less data
- What are my interaction sites ?
 - C^α ? C^β ? both ?
- Data bias
 - Can I ever find a representative set of proteins
 - PDB is a set of proteins which have been crystallised

Problems of Principle

- Boltzmann statistics
 - is the protein data bank any ensemble ?
- Is this a potential of mean force ? Think of Na, Cl example
 - that is a valid PMF since we can average over the system
- Energy / Free energy
 - how real ?
- N^{exp} ? how should it be calculated ?
 - is the fraction of amino acid a good estimate ? No.
 - there are well known effects.. Examples



Boltzmann based scores: improvements / applications

- collect data separately for $(i, i+2)$, $(i, i+3)$, ...
 - problems with sparse (missing) data
- collect data on angles
- collect data from different atoms
- collect protein – small molecule data

Are these functions useful ?

- not perfect, not much good for simulation
- we can take any coordinates and calculate a score
 - directly reflects how likely the coordinates are
- threading (coming soon)

Parameterising summary

- Inventing a score function / force field needs parameters
- totally invented (Crippen, Kuntz, ...)
- optimisation / systematic search
- statistics + Boltzmann distribution

Summary of low-resolution force fields

Properties

- do we always need a physical basis ?
- do we need physical score (energy)

Questions

- pick interaction sites
- pick interaction functions / tables

What is your application ?

- simulation
 - reproducing a physical phenomenon (folding, binding)
- scoring coordinates

Next

- even less physical