# Water models / solvation

## Biggest effects of water

- electrostatic
- dynamic

## Model types

- explicit
- implicit

Dynamic effects of water...

Andrew Torda, June 2018 strukt & sim 24/10/2018 [1]

# **Dynamic effects of water**

## one lonely moving particle



- initial velocity  $\dot{x}_t$
- future velocity easy  $\dot{x}_{t+\delta t} = \dot{x}_t$
- energy? constant  $\frac{m \dot{x}^2}{2}$

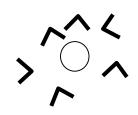
two particles? interacting?

- future velocity a bit more difficult
- easily bounded cannot be more than  $\frac{m_1\dot{x}_1^2 + m_2\dot{x}_2^2}{2}$

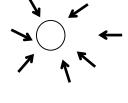
one particle in water...

# Velocities of particles in water

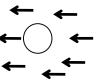
#### Lots of random interactions



A small acceleration?



A big acceleration?

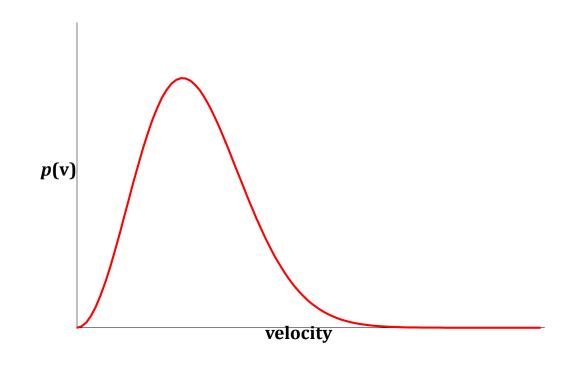


A probability distribution

+

how does  $\dot{x}_t$  tell us about  $\dot{x}_{t+\delta t}$ ?

much less



# **Modelling dynamic effects**

## Summary

- solvent will add fluctuations
- particles forget their velocity faster

#### Can this be modelled?

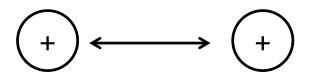
- yes (in molecular dynamics simulations)
- not really a force field / energy topic
- add random fluctuations to velocities
- can be made to look like water

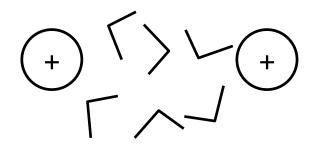
#### **Electrostatic effects of water**

#### water molecules

- not charged
- polar

Interaction between charges very different if water in between



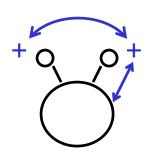


details soon

# **Explicit water**

#### Earlier descriptions of proteins

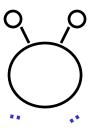
- a set of connected atoms
- extend to include water
- What does water look like?



- flexible angle
- stretchy bond
- charges

What else has it got?

- think about electron pairs on "O"
- what is really important?



# Important features of a water model

Do we care about water internal dynamics? (bonds and angle)

- usually not
  - make bonds rigid
  - make angle rigid
    - treat as a bond



#### **Dimensions**

- protons are really small
- does water geometry matter?
  - usually not



most important

Final result..





# SPC - A useful explicit water model

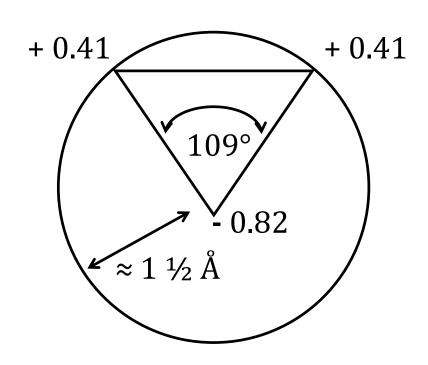
- 3 charges
- 1 Lennard-Jones radius
- 3 masses why?
  - only for molecular dynamics
- 3 bonds (completely rigid)
- Name "SPC", simple point charge

#### What can it do?

- diffusion, density, compressibility, heat capacity
- dielectric constant
- solvation energies?

#### Perfect? No

• add polarisation, offset charge from mass, ...



# Explicit water + protein

#### Protein-water interactions

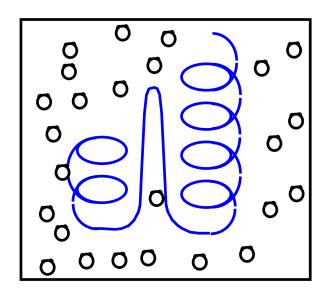
- via charge
- via Lennard-Jones term  $(r^{-12} \text{ and } r^{-6})$

## Elegant / Simple - automatically incorporates

- dynamic effects
- electrostatics

#### **Problems**

- very expensive
- typical simulation 10<sup>3</sup> protein atoms
- 10<sup>4</sup> solvent atoms



# worst case for proteins + water

Imagine a world with no cutoffs for interactions

- scales as  $O(n^2)$
- adding water gives 5 or 10 times as many atoms
- takes 25 or 100 times as much CPU time

#### Even worse

proteins move more slowly in water (viscosity)

#### What to do?

look for cheaper model

## Cheaper water models

Do we really need dynamic effects of water?

- maybe not
  - only want energies
  - only care about structures

or

model with a random force

Then look for model which gets most essential aspects of water

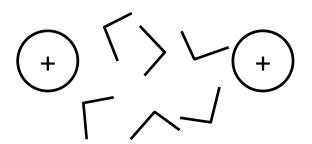
- electrostatics
  - distance-dependent dielectric
  - reaction field
  - surface area methods

## Distance-dependent idea

$$U(r_{ij}) = \left(\frac{1}{4\pi\epsilon_0}\right) \frac{q_i q_j}{r_{ij}}$$

$$= \frac{q_i q_j}{q_i q_j}$$

With solvent, 
$$U(r_{ij})$$
 changes less than  $\frac{q_i q_j}{D r_{ij}}$ 



#### Net effect?

- water is very polar and tends to orient itself around charges
- as if the water "screened" the charges (makes them smaller)

# Distance-dependent dielectric implementation

Invent approximation  $D_{eff} = r_{ij}$  then

$$U(r_{ij}) \approx \frac{q_i q_j}{D_{eff} r_{ij}} \approx \frac{q_i q_j}{r_{ij}^2}$$

Is this physics?

• no

Does it work?

- a bit (ugly)
- little real physical basis
- water does not behave so simply
- fundamental problem...

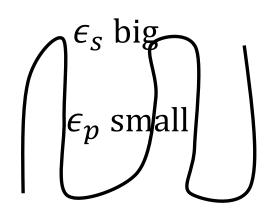
## Fundamental problem with distance-dependent D

If we rely on distance-dependent dielectric constant

• assume one 'fix' works everywhere (not true)

Think of formula 
$$U(r_{ij}) = \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}$$

Model will differ on big and small proteins

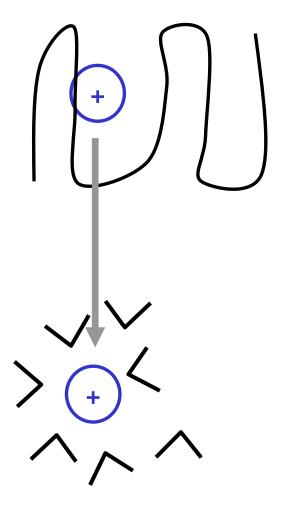


## Reaction field idea

## Different problem to before

- charge in a protein (lots of neutral CH groups)
  - not much happens
- particle in water
  - what does the water do?
  - tends to orient
  - lots of  $q^+q^-$  interactions
    - much better energy
  - is this like a force?
    - yes, think  $\frac{-dU}{dr}$

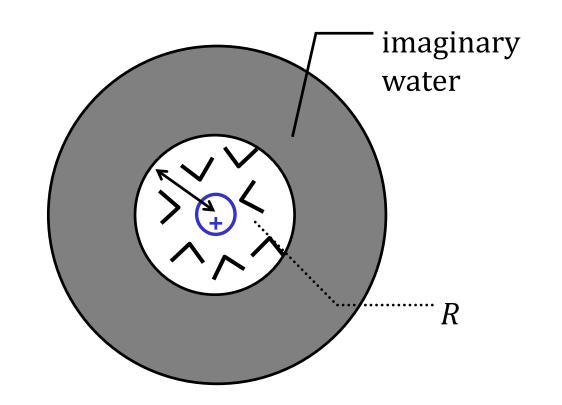
Can this be modelled?

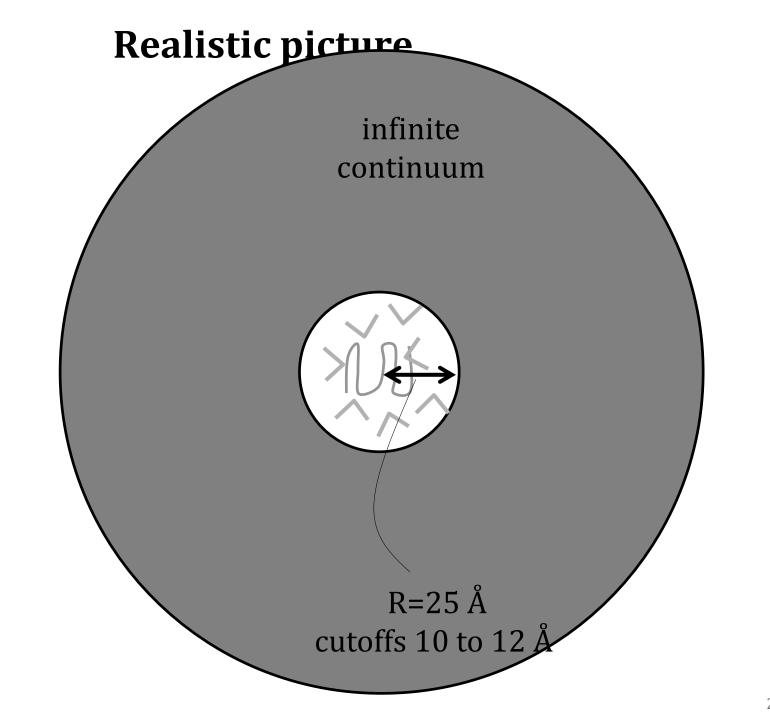


# interaction with imaginary solvent

Think of particle interacting with distant water molecules

- our charge interacts with them all but
  - if they are far away (big R) less important
  - depends on dielectric constant
  - inside white region  $\varepsilon_r$  and
  - grey region  $\varepsilon_s$
- within white region
  - treat atoms with a correction
- grey region
  - treat as continuum





# Reaction field / image charge formula

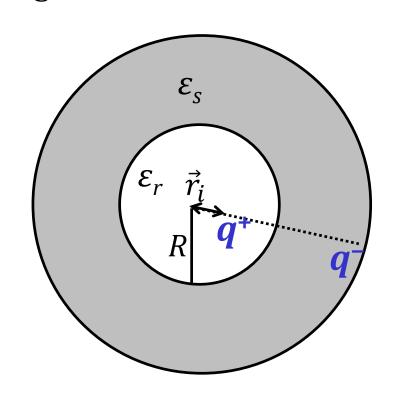
As if we interact with an "image" charge

• size 
$$q_{im} = -\frac{\epsilon_s - \epsilon_r}{\epsilon_s + \epsilon_r} \frac{q_i R}{r_i}$$

- location  $\left(\frac{R}{r_i}\right)^2 \vec{r_i}$
- near middle
  - $R \gg r_i$
  - image far away
- near boundary
  - imaginary solvent important
  - strong (favourable) interaction

## Important result

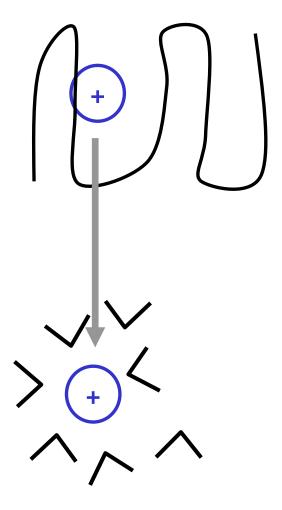
- we have modelled the happiness of a charge in solution
- charges happiest on outside of protein



# **Entertainment - why is this cheating?**

#### Newtons 3<sup>rd</sup> law

- there is a force on the +
- what is broken



# Simpler ways to model solvent

## Problem with real physics

- if you use this model, you are obliged to use
  - real charges, real coordinates...
- parameters not perfect
- hard to rationalise repairs

#### Many effects simultaneously

- charges interacting with water dipoles
- loss of water water interactions
- change of solvent entropy
- change of solute entropy?

#### Different approach

less rigorous models

# Basis of quick water model

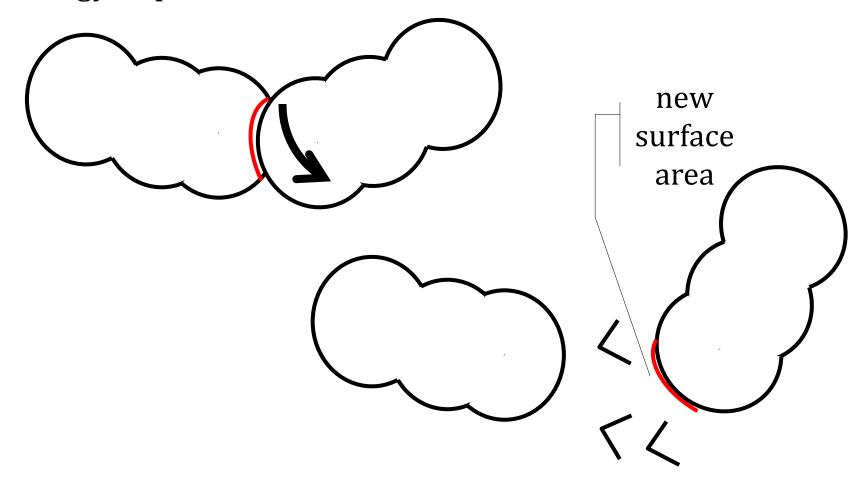
## Philosophy

- I can not model water properly
- find a very general way to incorporate effects
- Water makes some atoms happy
- Others do not care too much
- Find some very general way to include water effects
  - whether they are favourable / unfavourable
- what is easiest way to think about water influence?

## Atomic surface area

## Simple model

• for each atom, energy depends on surface area



# **Formalising SASA model**

- Solvent accessible surface area (SASA)
- for every atom, i  $G_i^{solv}(\vec{r}_i) = \gamma_i A_i(\vec{r}_i)$
- *G* because we no longer have a pure potential energy
- $G_i^{solv}(\vec{r}_i)$  because the energy term depends on coordinates
- $\gamma_i$  is a specific parameter for each kind of atom
  - for O, N will be negative
  - for CH, CH<sub>2</sub>, CH<sub>3</sub> will be positive or near zero
- area, *A*<sub>i</sub>, has to be calculated

#### **Problems**

- *A<sub>i</sub>* is difficult to calculate
  - use approximation
- $\gamma_i$  not easy to estimate

## **Example SASA calculation**

- classical atomistic force field
- distance-dependent dielectric
- two  $\gamma_i$  parameters,  $\gamma_{C,S} = 0.012$  and  $\gamma_{O,N} = -0.060$  kcal mol<sup>-1</sup>

#### Results

- better than in vacuo
  - deviation from known structure during simulation
  - not too many H-bonds formed
  - radius of gyration ? (how big is protein)
- why do they appear OK? why only two  $\gamma_i$ ?
  - not tested in detail
  - worst problems fixed

#### context

#### Who uses what?

- MD simulations
  - explicit water (very common)
  - reaction field
  - more complicated (long range periodicity)
- Drug design
  - occasionally do full MD simulations / free energy estimations /  $\lambda$  perturbation
  - fast screening
    - crude approximations

## summary

- Have not discussed dynamic effects of water
- Explicit water is best, but very expensive
- distance-dependent dielectric +
  - SASA style models
  - complementary
- many variations
  - surface accessible volume
  - more  $\gamma_i$  parameters
  - add in reaction field for better long range electrostatics
- changes and flaws in one parameter are hidden by others

# Coarse grain models (continuous) ... potentials of mean force

#### So far?

- very detailed models
  - 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  $\alpha$ -helical or  $\beta$ -sheet ?

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
- Score functions directly from structural data
- 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<sup>-9</sup> to 10<sup>-6</sup> 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

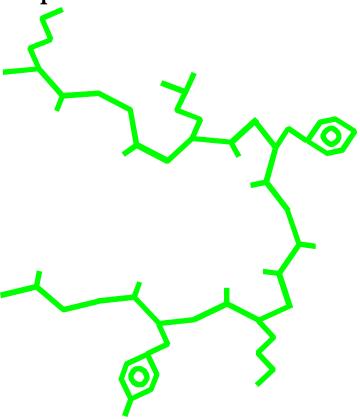
#### What will I need?

- 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?

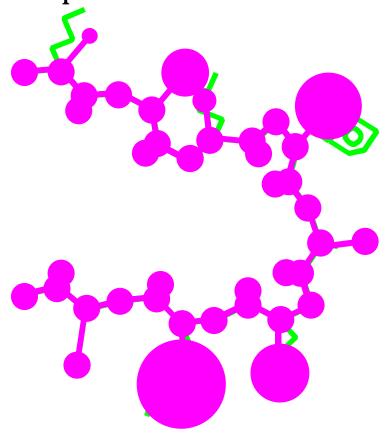
• decide on number of sites per residue



# **General implementation (easiest)**

How do we represent a protein?

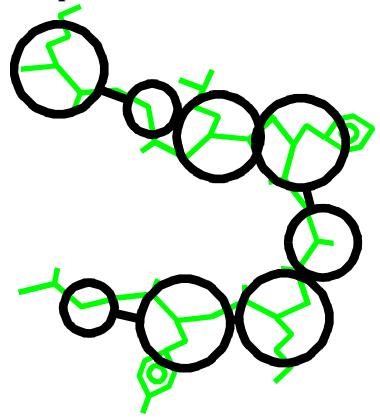
• decide on number of sites per residue



## **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

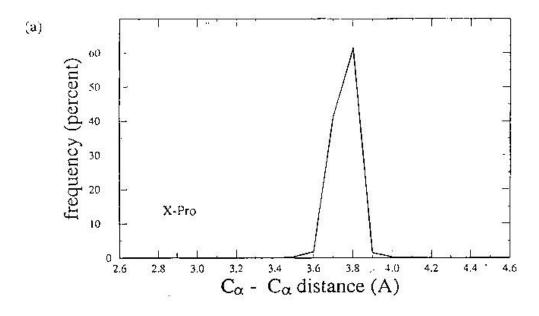
step through some examples from literature

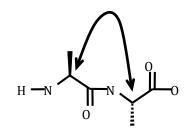
#### Common features

- 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  $C\alpha$  to  $C\alpha$  distances

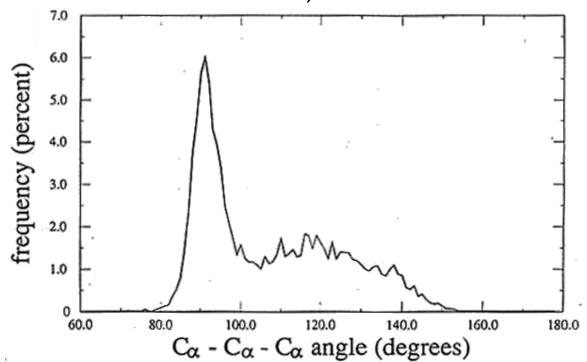


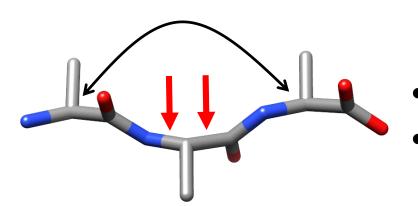


## Conclusion is easy

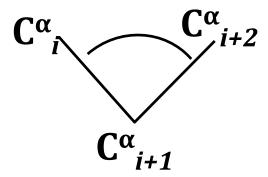
- any model should fix  $C_{i,i+1}^{\alpha}$  distances at 3.8 Å
- what other properties do we know?

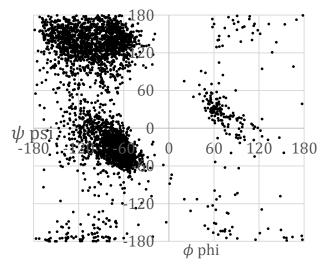
# $C_{i,i+2}^{\alpha}$ distance / angle





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





## First simple model

n residues, n interaction sites i, i+1 restrained ( $C^{\beta}$  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  $c_{ii}$  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 intented)
- physical pathways
- subtleties of sequence features (simplicity of  $c_{ij}$  matrix)

# Improvements to simple model

#### Aim

biggest improvement for least complication

#### **Possibilities**

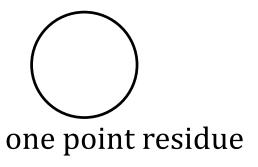
- more points per residue
- more complicated  $c_{ii}$  matrix... (more types of interactions)
- an example weakness

### Important structural features of proteins

- all proteins have hydrogen bonds at backbone
- proteins differ in their sidechain interactions...

# more complicated interactions

sidechain packing

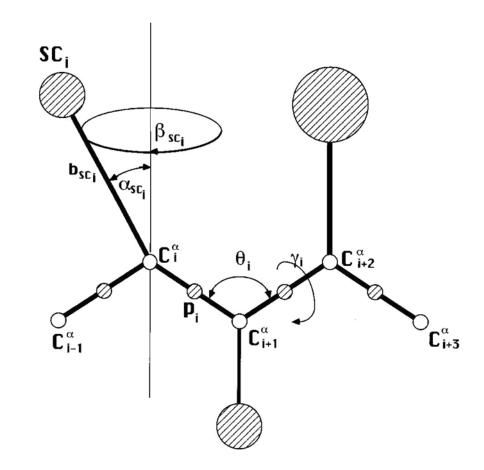


3 points per residue

## Scheraga model

## 3 points per residue

- 2 for interactions
  - p<sub>i</sub> is peptide bond centre
  - SC<sub>i</sub> is sidechain
- 1 for geometry
  - Cα
- $C^{\alpha}$   $C^{\alpha}$  fixed at 3.8 Å

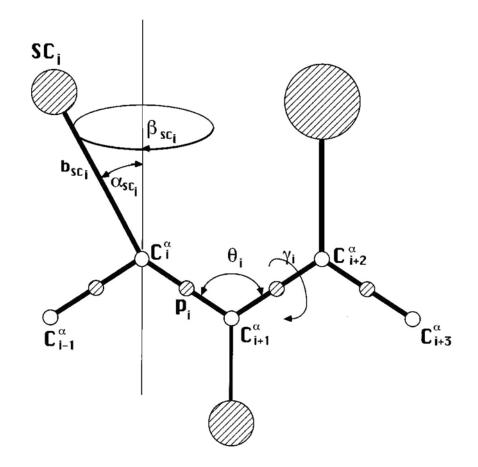


Do interaction sites correspond to atoms?

## Terms in Scheraga model

## Total quasi energy =

- side-chain to side-chain
- side-chain to peptide
- peptide to peptide
- torsion angle γ
- bending of  $\theta$
- ...
  - bending  $\alpha_{sc}$



# angle between $C^{\alpha}$ sites

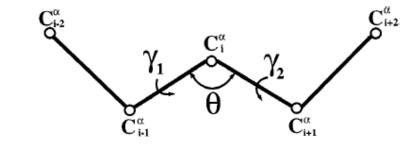
## Cunning approach

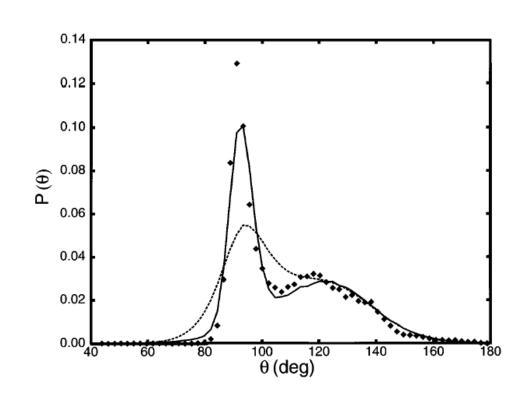
- look at  $\theta$  distribution
- model with Gaussians

then say

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

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





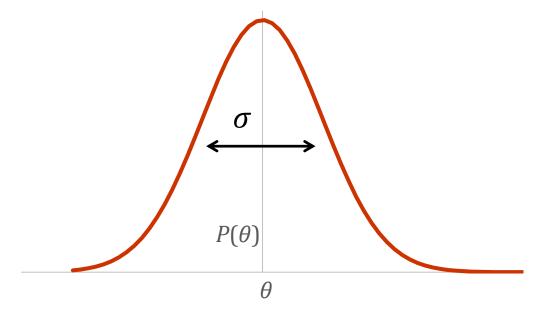
### Gaussian reminder

- get  $\mu$  and  $\sigma$  from fitting
- angle  $\theta$  depends on fitting

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

How would forces work?

- express  $\theta$  in terms of coordinates r
- say  $U(\theta)^{bend} = -RT \ln P(\theta)$
- take  $\frac{dU}{d\theta} \frac{\partial \theta}{\partial \vec{r}}$



## pseudo torsion term

Like atomic torsion  $U(\gamma_i) = a_i \cos n\gamma_i + 1 + b_i \sin n\gamma_i + 1$ 

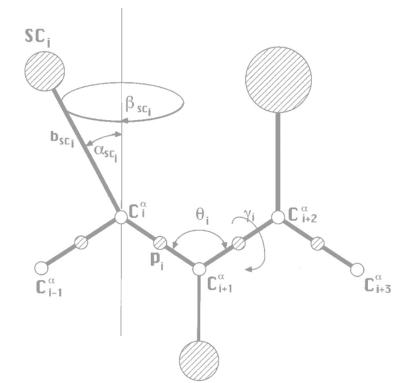
• *n* varies from 3 to 6 depending on types i + 1, i + 2 (numbering from picture)

Three kinds of pair

- gly
- pro
- others

#### Net result?

- residues will be positioned so as to populate correct parts of ramachandran plot
- this model will reproduce  $\alpha$ -helix and  $\beta$ -sheets



# side-chain peptide

### Not so important

- mostly repulsive  $U^{sc-p}(r_{sc-p}) = kr_{sc-p}^{-6}$
- *k* is positive, so energy goes up as particles approach

#### side chain interactions

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

but, consider all the  $\sigma$  and  $\varepsilon$ 

#### Main result

- some side chains like each other (big  $\varepsilon$ )
- some pairs can be entirely repulsive (small  $\varepsilon$  big  $\sigma$ )
- some not important (small  $\varepsilon$  small  $\sigma$ )

## more complications

#### Real work used

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

as
nicht für Klausul

#### 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 modelling 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 represents 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
- $\varepsilon$  and  $\sigma$  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?
- $\varepsilon$  and  $\sigma$ ?

# Approaches to parameterisation

#### General methods

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

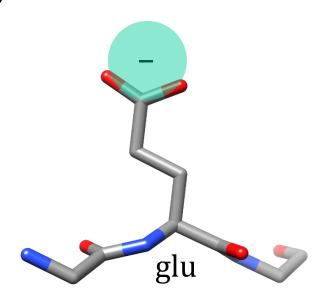
# From detailed to coarse grain

#### 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

- size of charge obvious
- location ?
  - not easy
  - does this let us include polarity? No.
- is this the right way to think about it ?...



## Averaging over details is not easy

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 / optimisation

```
for (parameter = small; parameter < big; parameter++)
    measure happiness</pre>
```

Define happiness - what do you want?

- density at equilibrium
- free energy change of some process
- distance of average protein structure from X-ray

• ...

### cost function

#### For your definition of happiness

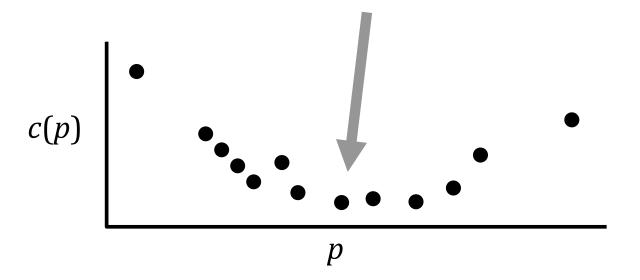
- some measured observable  $\mathcal{A}_{obs}$ 
  - density, dielectric constant, diffusion constant, ...

From simulation with parameter *p* 

- simulate and get  $A_p$
- unhappiness (cost) is a function of p, so we have c(p)  $c(p) = \left| \mathcal{A}_{obs} \mathcal{A}_{p} \right|$

or maybe 
$$c(p) = (\mathcal{A}_{obs} - \mathcal{A}_p)^2$$

concrete example..



- each point is result from a simulation
- noise / inaccuracy, not symmetric / linear

Example: parameter 
$$p$$
 is  $\sigma$  in  $U(r_{ij}) = 4\varepsilon_{ij} \left( \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{-12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{-6} \right)$  we would be adjusting the size of particles

# parameters optimisation - boring? easy?

You would not choose *p* values randomly or by systematic search

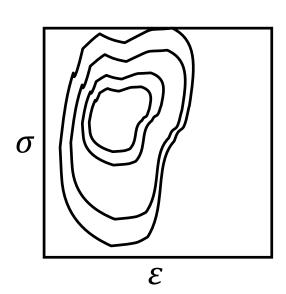
• (use a classic optimisation method)

Is this too easy and dull?

• what you probably have is several parameters  $c(p_1, p_2)$ 

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

• measure the error/cost in 2D space



## mapping parameter space

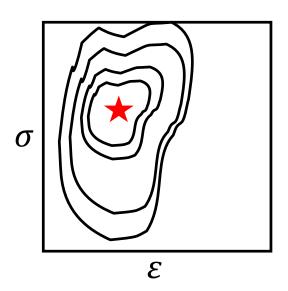
#### What does this tell us?

- find best  $\varepsilon$  and  $\sigma$
- see that  $\varepsilon$  is critical,  $\sigma$  less so

#### Practical implementation

- systematic search? Inefficient
- automate the optimisation

Problems...



# **Problems with parameterisation**

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

### Optimising for some properties

- you optimize for density
  - diffusion, free energy changes ....
    - all broken

#### Generalisation / recall

- you optimise based on 10 proteins
  - test of 11<sup>th</sup> bad results (too small training set)

### Different kind of score function

Change of style...

- questions on coarse-graining?
- why is entropy an issue? (numbers of particles / states)
- from nice ideas to dumb empiricism

#### Potentials of mean force

Potential of mean force ... knowledge based score functions

- very general
- history from atomistic simulations

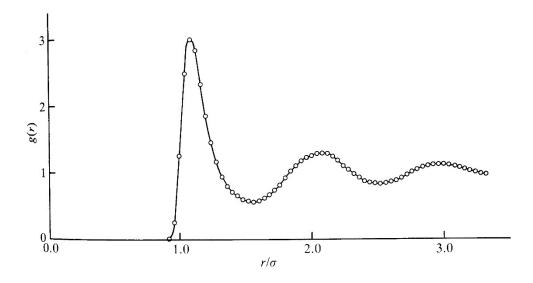
Basic idea .. easy

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

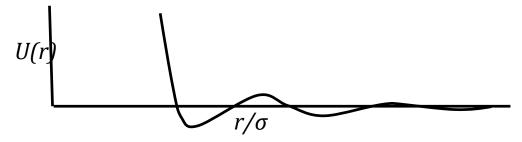
## Intuitive version of potential of mean force

### Radial distribution function g(r)

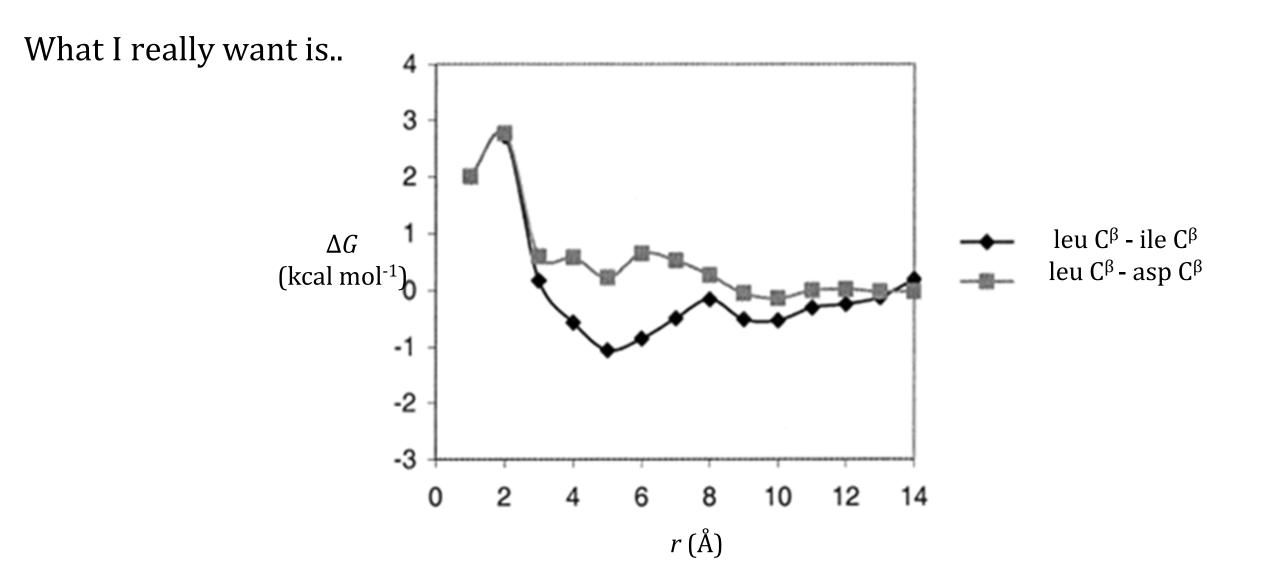
• probability of finding a neighbour at a certain distance



What does this suggest about energy?



## Goal



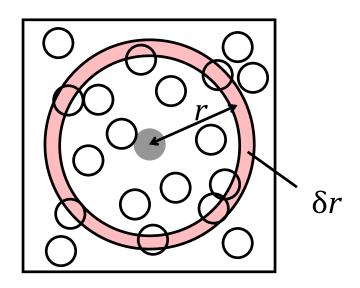
## **Radial distribution function**

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

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

- *N* particles
- V volume
- Calculating it?
  - define a shell thickness ( $\delta r$ )
  - around each particle
    - at each distance, count neighbours within shell

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



# Rationale for potentials of mean force

For state i compared to some reference x

$$\frac{p_i}{p_{\chi}} = \frac{e^{\frac{-E_i}{kT}}}{e^{\frac{-E_{\chi}}{kT}}} = e^{\frac{E_{\chi} - E_i}{kT}}$$

$$\ln \frac{p_i}{p_\chi} = \frac{E_\chi - E_i}{kT}$$

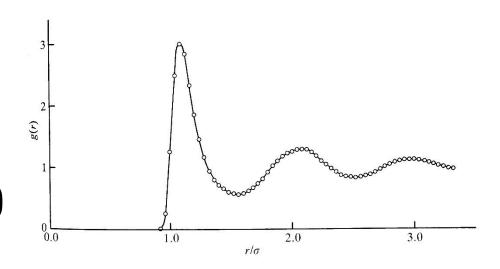
$$\Delta E = kT \ln \frac{p_i}{p_x}$$

### 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
  - strength of interactions / order

Relate this back to energy



# Energy from g(r)

From statistical mechanics 
$$g(r) = e^{\frac{-\Delta G}{kT}}$$

- not quite textbook formulation normally describe in terms of work
- what is the  $\Delta$  in  $\Delta G$ ?
  - the free energy change going from infinite separation to some distance r

## How would we get g(r)?

- experiment? sometimes
- simulation easy simulate at high resolution
- soon protein data bank

#### Assumptions

our system is at equilibrium

# Generalising ideas of potential of mean force

What else can we do?

• think of more interesting system (H<sub>2</sub>0)

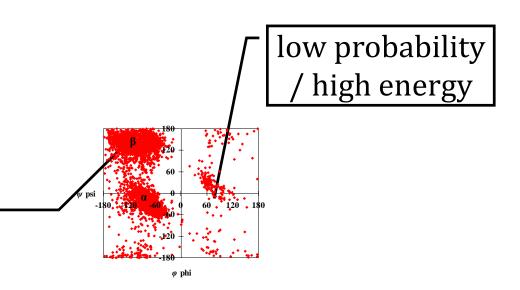
Would we express our function in terms of O? H?

- both valid
- could consider work done bringing an O to O, O to H, H to H
  - for fun on next page

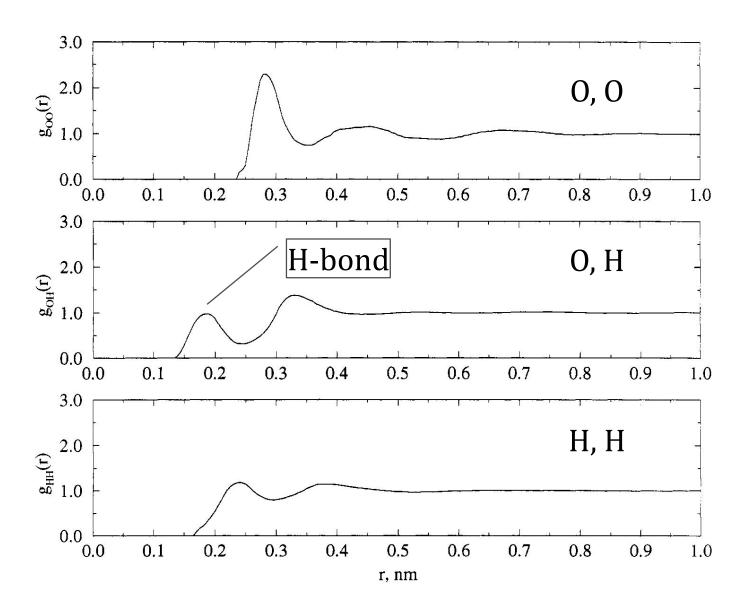
### More general..

- are we limited to distances? No
- example ramachandran plot
- other atoms? ...

high probability / low energy



## radial distribution function (water)



## 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

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

 $N_{AB}^{obs}$  how many times do we see

- particles of types A and B
- distance r given some range  $\delta r$

 $N_{AB}^{exp}$  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<sup>-</sup> and Na<sup>+</sup> in water like to interact (distance  $r^0$ )
- $N_{AB}^{obs}$  is higher than random particles
- $W_{\text{CINa}}(r)$  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

$$N_{AB}^{exp} = N_{all}X_{Na}X_{Cl}$$
 (use mole fractions)

• use this idea to build a protein force field / score function

### **Protein score function**

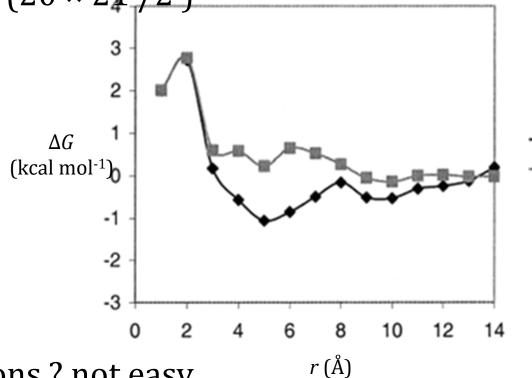
## Arbitrarily

- define interaction sites as one per residue
  - maybe at  $C^{\alpha}$  or  $C^{\beta}$
- 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 Nobs
  - how many pairs of type gly-gly, gly-ala, gly-X, X-Y... are there?
    - gives me *N*<sup>exp</sup>
  - repeat for 5 Å, 6 Å, ...
- resulting score function...

## final score function

For every type of interaction AB  $(20 \times 21/2)$ 

• 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

leu  $C^{\beta}$  - ile  $C^{\beta}$  leu  $C^{\beta}$  - asp  $C^{\beta}$ 

# Applying knowledge-based score function

## Take your protein

- for every pair of residues
  - calculate  $C^{\beta}$   $C^{\beta}$  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  $\left(\frac{dU}{dr}\right)$
- it is not necessarily defined for all coordinates

### Practical Problems Boltzmann score functions

Do we have enough data?

how common are Asp-Asp pairs at short distance?

How should we pick distance ranges? How far should we look?  $(r_{AB})$ ?

What are my interaction sites?

•  $C^{\alpha}$ ?  $C^{\beta}$ ? both?

#### Data bias

- Can I ever find a representative set of proteins?
  - PDB is a set of proteins which have been crystallised

### Reminder

- we want low-resolution score functions
- if we work in a Boltzmann framework, we work with real energies
- everything ends up as  $\frac{p_i}{p_j} = e^{-\frac{\Delta E}{RT}}$  or here  $\Delta E = -RT \ln \frac{p_i}{p_j}$  or  $\Delta E = -RT \ln \frac{N_{obs}}{N_{exp}}$
- we are comparing against what you expect from random events without interactions  $\boldsymbol{p}_i$
- work with kJ mol<sup>-1</sup>, we can
  - make real energetic predictions (kinetics, equilibria)
  - combine with other energy terms

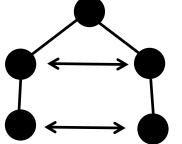
# **Problems of Principle**

#### **Boltzmann statistics**

- is the protein data bank a set of structures at equilibrium? 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?

### *Nexp* ? how should it be calculated ?

- is the fraction of amino acid a good estimate? No.
- there are well known effects.. Examples



$$i,i+2$$

*i,i*+4 very different statistics

# 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 / fold recognition / model quality

# **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

#### **Parameterisation**

Averaging, optimisation, potentials of mean force

Next – less physical