Analysis and comparison

Stories

- 1. quality
- 2. surfaces
- 3. Comparing structures

Quality

Meaning?

- How good is the electron density ?
- How well are atoms placed ?



experimental issues

Crystal quality and size

- NaCl, sugar,.. crystallise in the kitchen
 - crystals large
 - soup \rightarrow ordered state, ΔG is favourable
- proteins
 - not so regularly shaped
 - ordered arrangement may not be much better than random orientations
 - which has better free energy ? entropy ?





nice crystals / bad crystals

You get a crystal – some disorder

- you see the average
- if the position of atoms varies the coordinates are
 - smeared not well determined

Result

- resolution not so good
- atoms are put in wrong places
 - sidechains fit to noise, water, ..

Judging the structure

Two sides

- 1. fit to experiment (in Biophys lectures) R and R_{free}
- 2. how good is the structure itself? (this topic)

What do people look at?

- energies ?
- geometries properties

Why do we not use energies (so much)?

Energies – not easy to use / assess

Two proteins with 100 residues

- 1. lots of big hydrophobic residues lots of van der Waals
- 2. a long protein

small core, interactions with water and ions

Difficult to compare energies



You give me a protein and energy calculation

• can I judge the coordinates ? Not easy

What can one look at?

typical properties of proteins

2wpz

Typical properties of proteins

- Ramachandran plots
- side-chain distributions
- clashes

Ramachandran outliers

Random sampling of protein data bank

- which are bad and which are OK?
- not every site is *α*-helix or *β*-sheet

• some example proteins

not all bad some small residues



happy coordinates

Why do I know they are happy ?



4wmx

happy Ramachandran plot

Each unusual residue was checked and OK

• gly and ser





A bad structure



bad Ramachandran

unlikely residues cannot be explained



15/11/2018 [12]



Where are the problems in bad structure?



Loops?



- often did not know where atoms are
- placed them not in most likely positions

Clashes

Best method to assess ? energies

Fastest method

• for each atom *i* we have a radius r_i (textbook)

44 N

224 0

his

ile 227 CG2

225 NZ phe 197 CB

phe 38 CE1 val 60 CG1

• for each pair of atoms calculate d_{ij}

if $d_{ij} < r_i + r_j$ complain

asn

lys

glu

• bad coordinates ...

too small (Å) 0.8 0.6 0.6 0.6 0.6

43 ND1

a clash

- not so dramatic
- 1/2 Å, small but important in energy

phe 38 CE1 val 60 CG1 0.6 Clashes are not so easy to judge • bad energy, but • geometry is very close to correct

5glw

sidechain rotamers

Torsion angle – energy model OK – not usually used Empirical approach

- visit high-resolution structures in PDB
- collect data on each side-chain angle make histograms

Look at coordinates

 for each sidechain angle decide on probability



rotamer modelling



Comparing good and bad examples

Both proteins (4wmx, 5glw)

- 2 Å resolution
- year 2017

Does it matter ?

- No. clash errors small $\approx \frac{1}{2}$ Å
- Yes. parts of backbone are fiction

Depends on application

- comparing with other proteins ? not important
- discussing ligand binding ? important

Where do problems come from ?

• Data – weak – where their software put atoms



summarise quality

Ramachandran plots

- physics torsion angles, Lennard-Jones, electrostatics
- we look at frequencies in protein data bank
 Clashes
- physics Lennard-Jones and electrostatics
- we look at hard radii

Rotamers

- physics torsion angles, Lennard-Jones, electrostatics
- we use frequencies from protein data bank

Can we justify this?

Good energy models or rough





r_{ii}

Statistics / counting (rotamers, backbone angles)

• what we see in the world reflects energies (Boltzmann relation)

$$p_i \propto \exp\left(\frac{-E_i}{kT}\right)$$

formula not for Klausur

What do we think of unhappy structures ?

Are they necessarily wrong ? ask again in 3 minutes



• maybe the side-chains have moved ?

- differences are small
- agreement with data is no worse

 with a bit of effort authors could have avoided this attention



• there are some terrible structures in PDB search for obsolete PDB (just for fun)

Surfaces

- not really a quality issue
- a property that quickly says if something is unusual

What do you expect?

- surface must be more charged and polar than the middle
- lots of -ve or +ve charges ? not so common
 acidic or basic proteins do exist
- charged regions ? Interaction with substrates ?
- very neutral will not be soluble

peroxidase





╋



1qgj

an acidic protein



Structure comparisons

- Why ? Function prediction evolution
- Sequence versus structure conservation..

Simple view of molecular evolution

mutate continuously

- mutations which are not lethal
 - may be passed on (fixed)
- if structure changes
 - protein probably will not function
 - not passed on



Result

 nature tests many sequences and keeps those that are compatible with structure

Structure determines function, but..

What is more informative

• sequence or structure similarity ?

	sequence	structure
	similarity	similarity
closely related	yes	yes
less similar	no	yes
not related	no	no

- look for sequence similarity most helpful
- structure similarity 2nd choice
 - relationships that one would miss

Sequence versus structure alignment

- Aim: why can we not use sequence alignment methods
- Sequence alignment reminder (more in summersemester)
- reminder

sequence alignment

Seq	ID	40.6	5 응	(1	03 /	25	54)	in	280	tota	1	including		gaj	ps			
	:	1		:	2	2	:		3	:		4	:		5		:	6
	:	0		:	C)	:		0	:		0	:		0		:	0
kkar	kkapviwvqqqqctgcsvsllnavhprikeilldvislefhptvmasegemalahmyeia																	
krpsvvylhnaectgcsesvlrtvdpyvdelildvismdyhetlmagaghaveea-l-he																		
:	:	1		:	2		:	3	3	:	4		:	5	5	:		
:	:	0		:	0		:	(כ	:	0		:	()	:		
	:	0		:	C)	:		0	:		1	:		1		:	1
	:	7		:	8	3	:		9	:		0	:		1		:	2
	:	0		:	C)	:		0	:		0	:		0		:	0
ekfr	ekfngnffllvegaiptakegrycivgeakahhhevtmmelirdlapkslatvavgtcsa																	
aikg-dfvcvieggipmgdggywgkvggrnmydicaevapkakaviaigtcat													cat					
0	:	:	0		:	0				:	0		:	1	L	:		1
6	:	:	7		:	8				:	9		:	(0	:		1
0	:	:	0		:	0				:	0		:	()	:		0
	:	1		:	1	-	:		1	:		1	:		1		:	1
	:	3		:	4	ł	:		5	:		6	:		7		:	8
	:	0		:	C)	:		0	:		0	:		0		:	0
yggi	ipaa	aegnv	rtg	sks	vrdf	fac	lek	iek]	Llvn	vpgcp	ph	pdwm	vgt	lva	aaw	shv	lnpt	ceh
yggv	7qaa	akpnr	otg	tvg	vnea	lgł	lg	vkai	in:	iagcp	pn	pmnf	vgt	.v	-vh	11t	k	
:	:	1		:	1		:	1	L	:		1	:	:		1		
:	:	2		:	3		:	4	1	:		5	:	:		6		
:	:	0		:	0		:	()	:		0		:		0		
	:	1		:	2	2	:		2	:		2			:		2	:
	:	9		:	C)	:		1	:		2			:		3	:
	:	0		:	C)	:		0	:		0			:		0	:
plpe	eldo	ldgr	11:	ffg	dnih	enc	py	ldky	ydnse	efaet	ft	kpg-		cł	cae	lgc	kgps	sty
gmpeldkqgrpvmffgetvhdncprlkhfeagefatsfgspeakkgyclyelgckgpdty																		
:		1	:		1	:		1		:	2	:		2		:	2	2
:		7	:		8	:		9		:	0	:		1		:	2	2
:		0	•		0			0		•	0			0		•	C)



Sequence alignment steps

 $c d e f g h i k \dots$

a

1

m

n

p

q

r

S

steps

- similarity score
- sum up possible paths
- find optimal path

First step –similarity

look up in a table (blosum matrix)^t/v
 how similar is

a to 1, a to m, a to n, .. c to 1, ..

Can one do this with structures?

Difficulty with structure alignments

- to build a score matrix, must compare 1 to 1', 2', ...
- 2' depends on 1', 3'
 - 1', 3' have not been aligned
- there is no obvious similarity measure comparing two sites in structure

Time for guesses / approximations







Sequence philosophy – structure alignments

If each part of a structure has a label, can compare labels

4 • say 1 is α , 2 is α , 3 is γ • similar labels in red structure 6 5 can build a score matrix 2 3 4 5 6 7 8 9 1 • fill with 1's and 0's 2 3 4 5 6 Could one use secondary structure? • would it work? 7 8 not well 9

Alignments based on secondary structure

Problems

- 1. alphabet is too small
- does not capture similarity
 - lots of alternative alignments of nearly equal score
- 2. requires regular structure







5lus

Labels on pieces of structure

Classic 2° not enough – better alphabet for structures

Break structure into fragments





2 proteins

- split into pieces
- compare fragments
- score and find best path
- will give an alignment

How does one compare fragments ?

· **7** ·

8 9

6

2 3

1

5

6

7

8

9

...

4 5

Comparing fragments - angles

Measure of similarity

Example – angles

- turn into a list of ϕ , ψ angles
 - \vec{p}_1 is $[\phi_1, \phi_2, ..., \phi_6, \psi_1, \psi_2, ..., \phi_6]$
 - get \vec{p}_2 for protein 2
 - calculate $d = |\vec{p}_1 \vec{p}_2|$
 - put $\frac{1}{d}$ into score matrix
 - if two fragments are similar, big positive value for similarity
- Why is it nice ?
- works on regular structure or strange structure



Comparing fragments - distances

For each fragment

- look at C^{α} in middle
- get list of distances to C_{i-3}^{α} , C_{i-2}^{α} , ...
- the fragment is set of distances \vec{d}_1
- I can compare this vector of distances for different fragments

There will be a set of characteristic distances for

• α -helical fragments, β -sheet, common turns, anything

Optimal alignment

- some measure of similarity
- move protein 2 on to protein 1?
 - will not work if proteins are not identical



structure alignments - no correct answer

Two very similar proteins

• align parts perfectly



or

 align the whole proteins less exactly ?

Arbitrary

 how many residues to align



Properties of structure alignments

- Much slower than sequence alignments
 - calculate fragments, angles, distances, ..
- no statistical basis (sequences use exchange frequencies)
- gap penalties trial and error
- no definition of optimal

Quantifying similarity

Full information – the two proteins

- similar overall shape
- differ in the middle
- must be evolutionarily related
- probably same function



• too hard

What we work with – one or two numbers

• give an idea of similarity

What do you want to tell me?

• typical distance between sites

What sites ?

- Serious crystallographers comparing nearly identical structures
 - all atoms
- Most literature comparisons
 - much less ...



What atoms ?



Atoms used

Most common choice

 C^{α} atoms

- 1. present in every residue
- 2. a set of α carbons nearly defines the shape of a protein

How to get a single number for comparing structures



root mean square differences

To characterise the spread in a set of numbers

• standard deviation $\sigma = \left(\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2\right)^{1/2}$ where \bar{x} is mean

To characterise the structural differences

$$rmsd = \left(\frac{1}{N}\sum_{i=1}^{N} |\vec{r_i} - \vec{r'_i}|^2\right)^{1/2}$$

 $\vec{r_i}$ is atom coordinates in first structure $\vec{r_i'}$ coordinates in second structure

• rms / *rmsd* / RMSD / r_{rms} = root mean square difference

Some assumptions

We have already done rotation and translation



We have a list of matching atoms (from the alignment)

coordinate *rmsd* is evil



many alternatives to rmsd

Sociologically important

- 1. TM-score (template modelling)
- scales distances depends on alignment length
- 2. GDT-score (global distance test)
- superimpose two structures and see how many residues can be superimposed with *rmsd* < 1 Å
- repeat with 2, 4, 8 Å
- get average
- what is the advantage ?

alternatives to rmsd - advantages

Why use these methods?

- values from 0 (very different) to 1 (identical)
- less size-dependent

Why not use these methods

- very protein specific
 - \bullet assume residues / C^{α} sites
 - not even good for nucleotides
 - no help for small molecules

Summarise

- Why energies are not often used
- What properties does one look at ?
- Physics vs. statistics
- Why structure alignments are difficult
- Different ways to quantify similarity