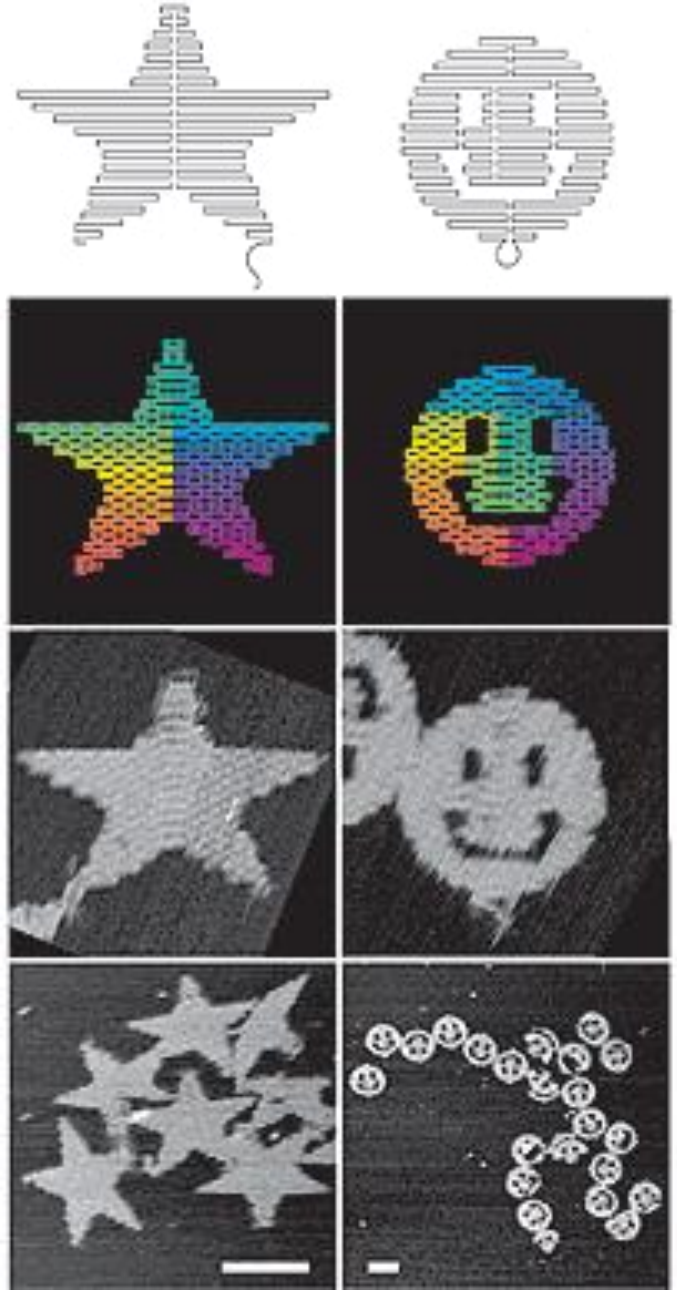
Nucleotide Design

Mission

- design large structures from DNA
- design smaller from RNA

Different to protein design

- conformations
- energies…



Rothemund, P.W.K., Nature 440, 297-302 (2006)

Andrew Torda, wintersemester 2016/17

Energies

True physics

- atoms interact with each other (electrostatics, Lennard-Jones, bonds..)
- works for proteins, nucleotides, old shoes, …

What happens here ?

- use approximations to catch most important effects

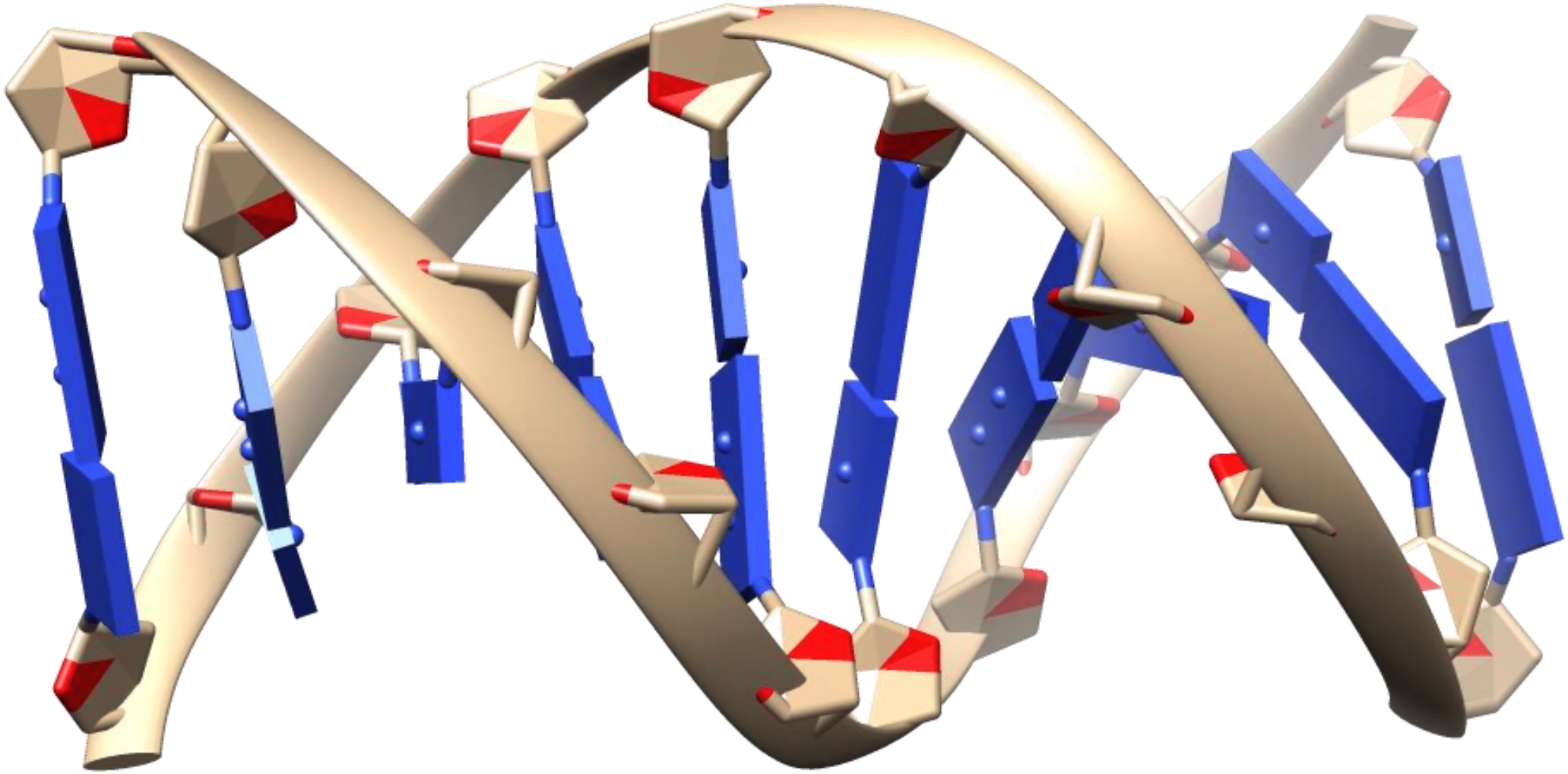
Protein

- approximations that capture the important physical effects
  - "fitting" to backbone, fitting with each other

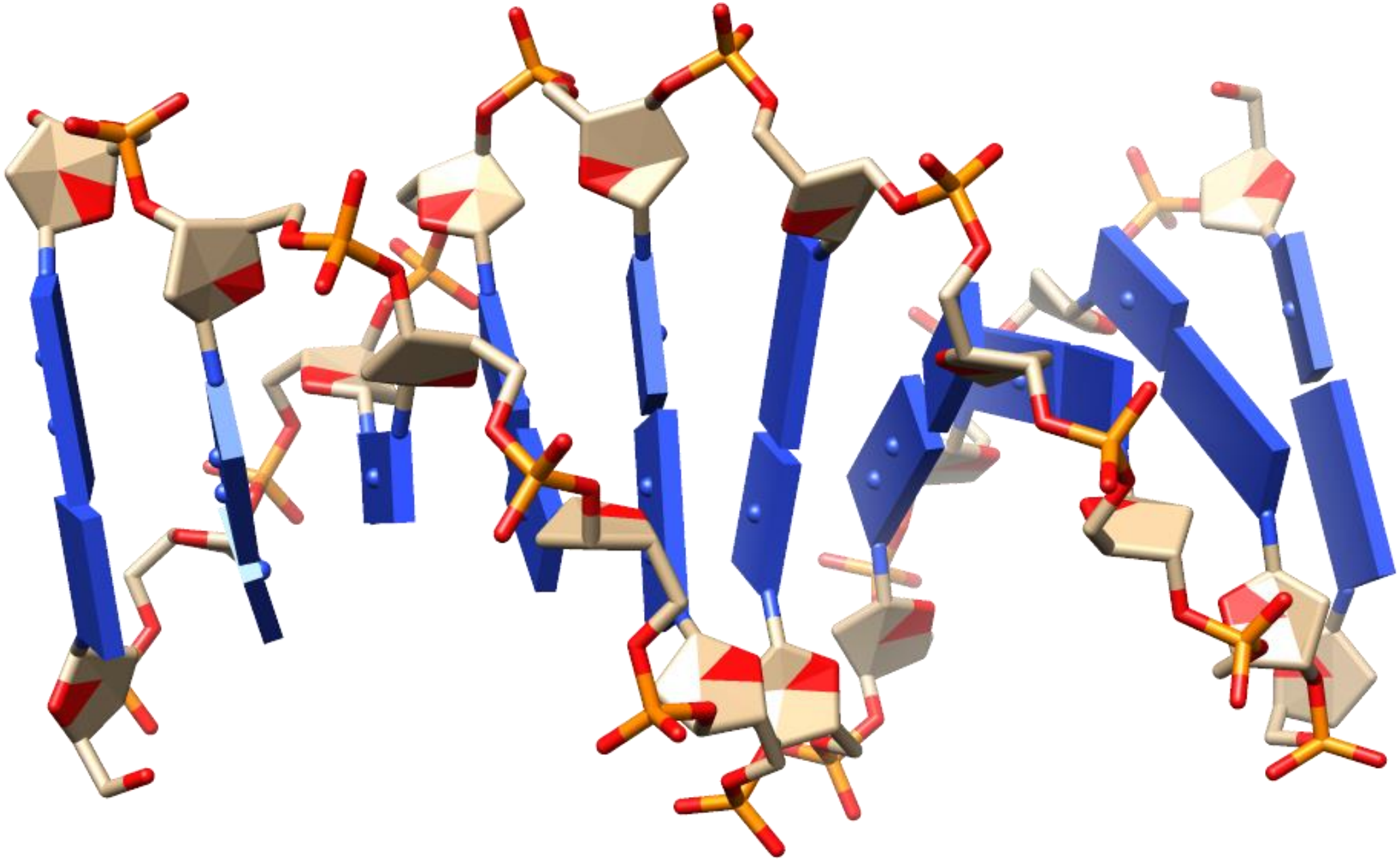
Nucleotides - what is important ?

- Hydrogen bonds and stacking - first H bonds

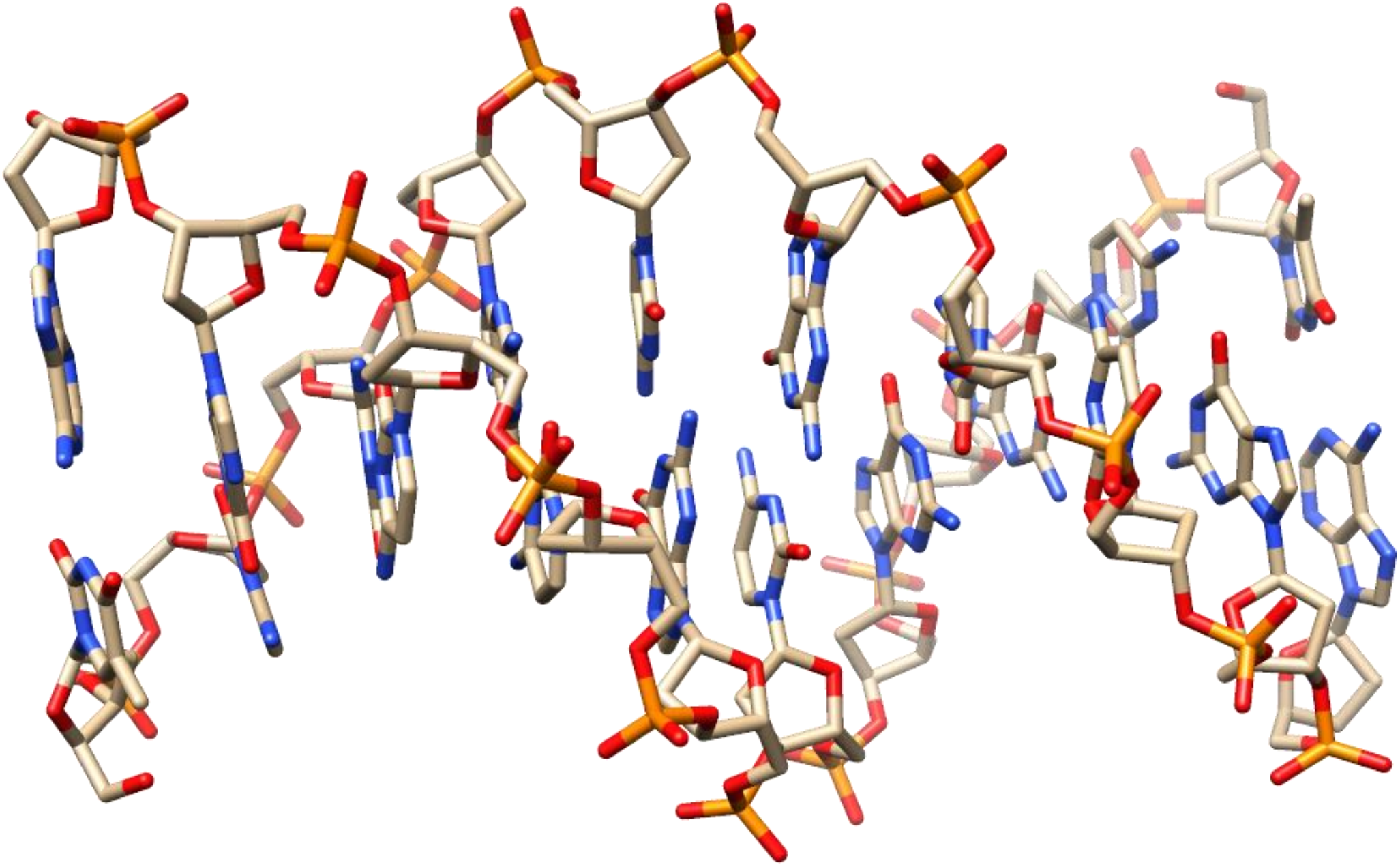
DNA very idealised



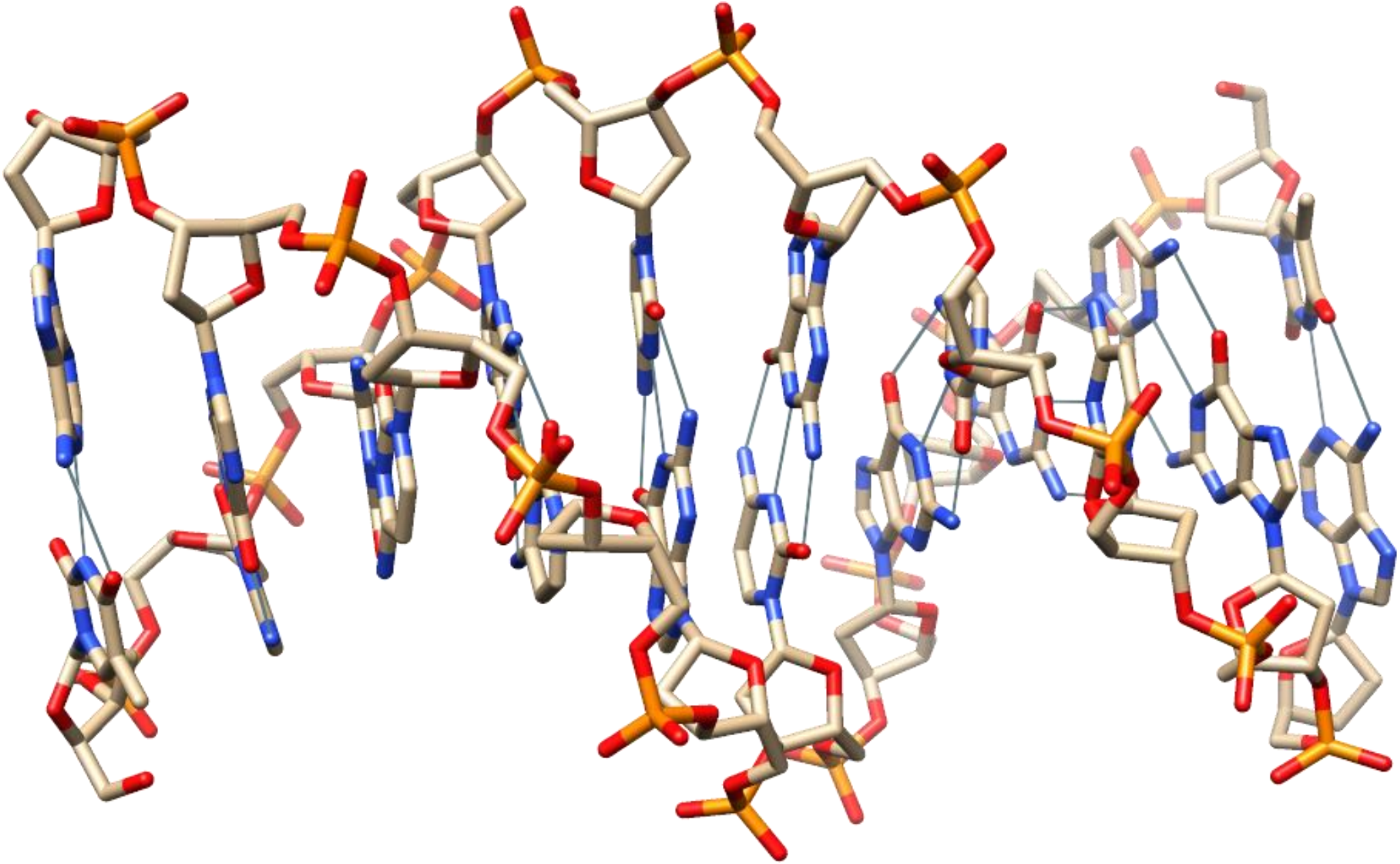
DNA backbone is not so smooth



DNA all atoms



DNA with Hydrogen bonds



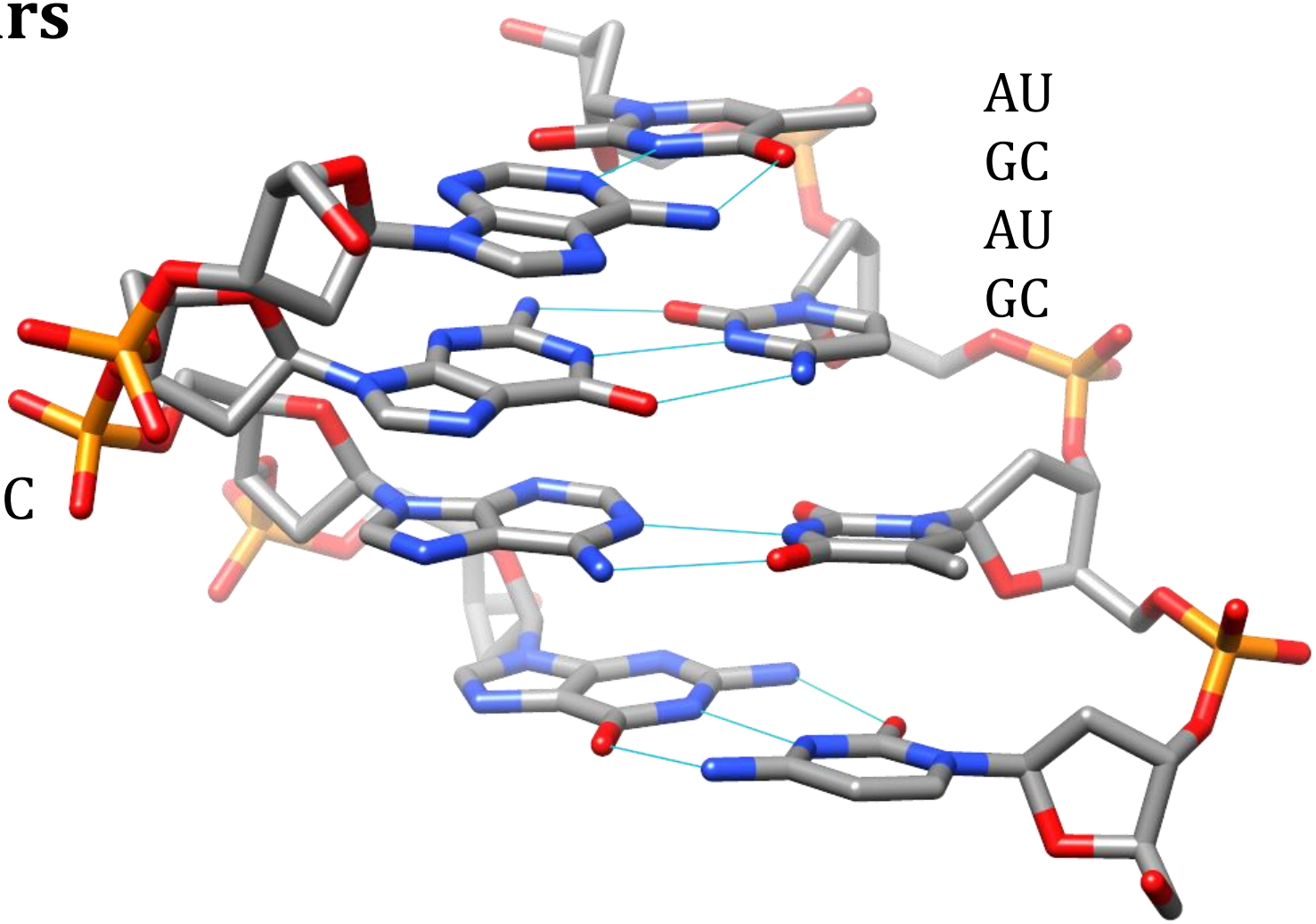
Energies – base pairs

Base pairing

- GC – 3 H bonds
- AU – 2 H bonds

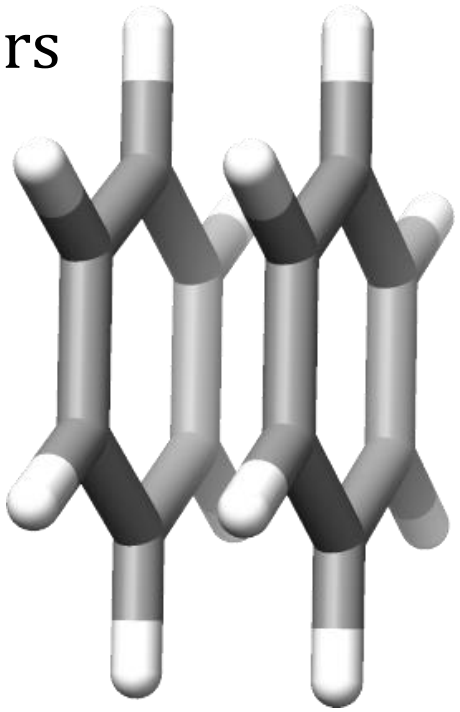
Sequence is happier with more GC

- not so simple (later)



H bonds and base pairing

- DNA philosophy – dominated by base pairing between two strands
- RNA –usually single stranded – folds up on itself, base pairs



Base pairing is very important

- try to form GC, AT pairs (DNA) or GC, AU pairs (RNA)

Is it the only important thing ?

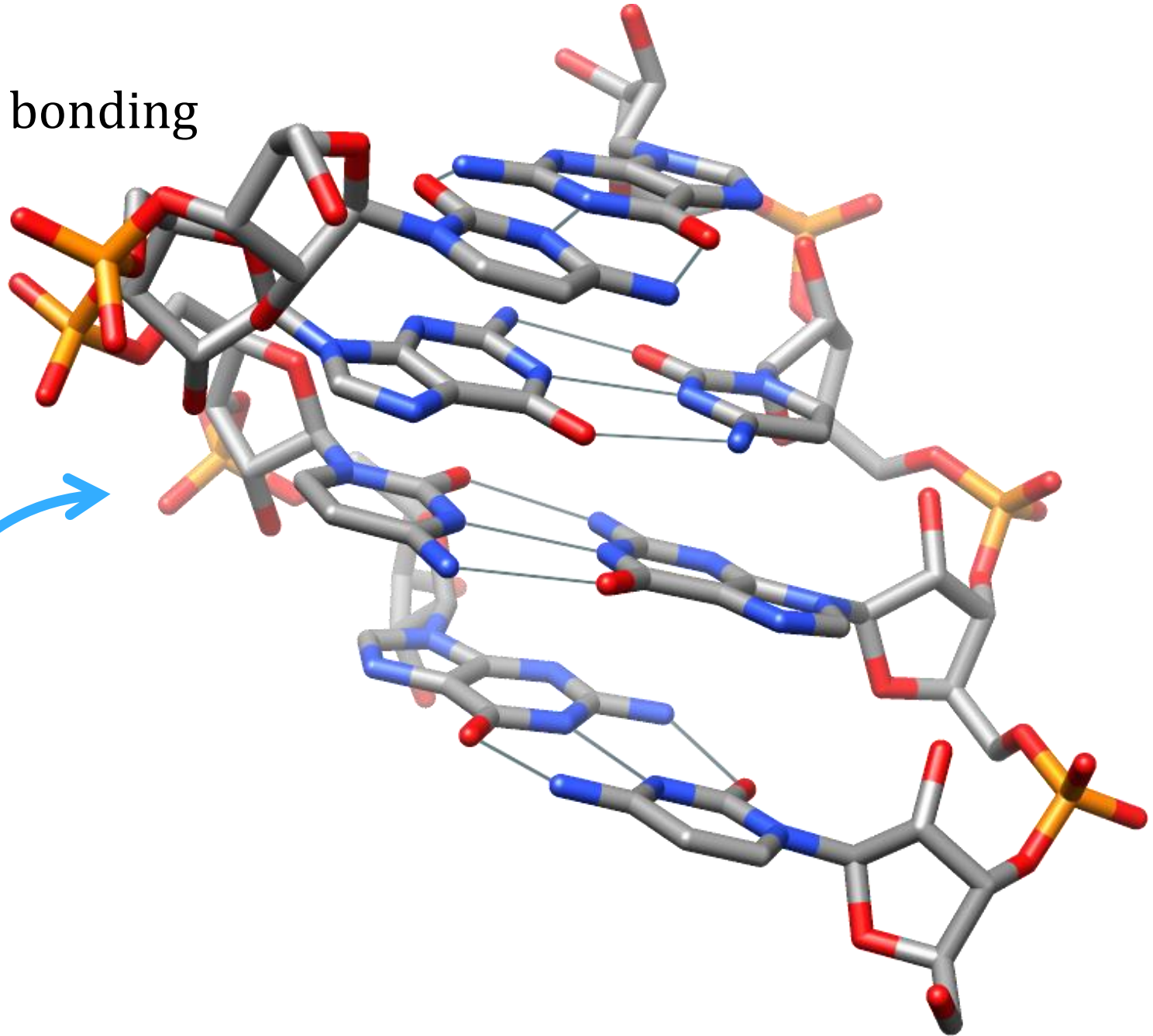
- aromatic ring stacking, π-stacking, base-stacking, …

GC GC

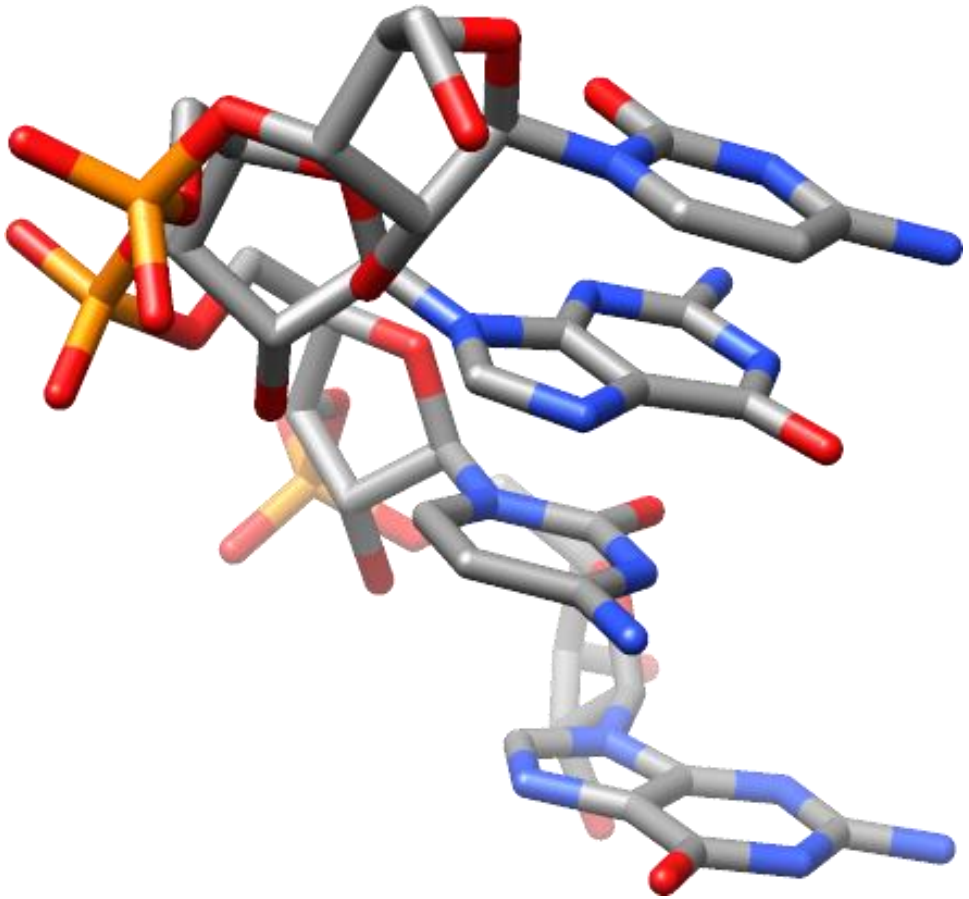
First think of hydrogen bonding

- then…

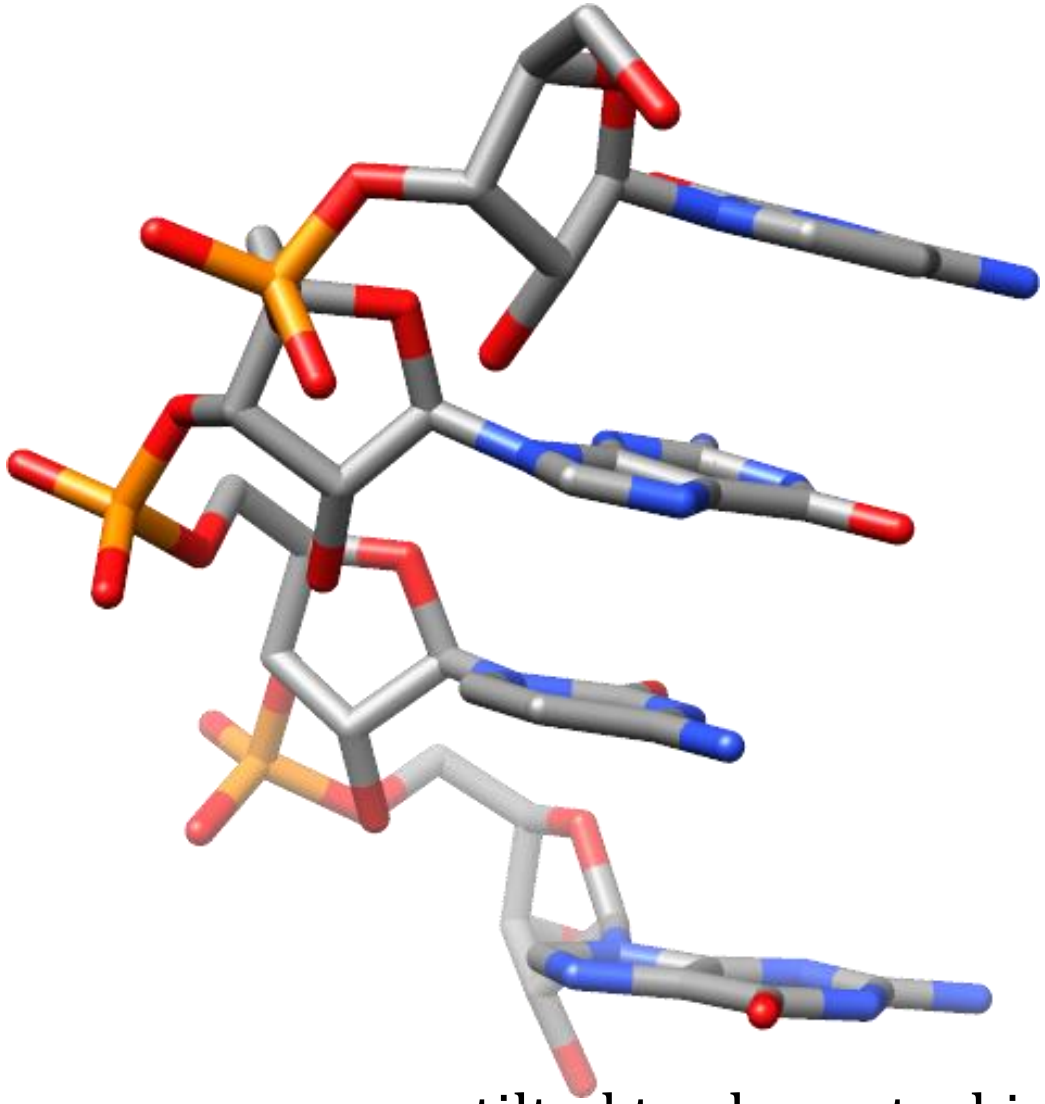
Now, look at just one strand…



Base stacking



as on previous slide



tilted to show stacking

Summarise energies

Just approximations – there are much better models for physics

Base-pairing

- important
- GC vs AU or AT

Stacking

- energetically favoured – structures are happy when they are regular and put bases on top of each other

Using energies

Literature (not physics)

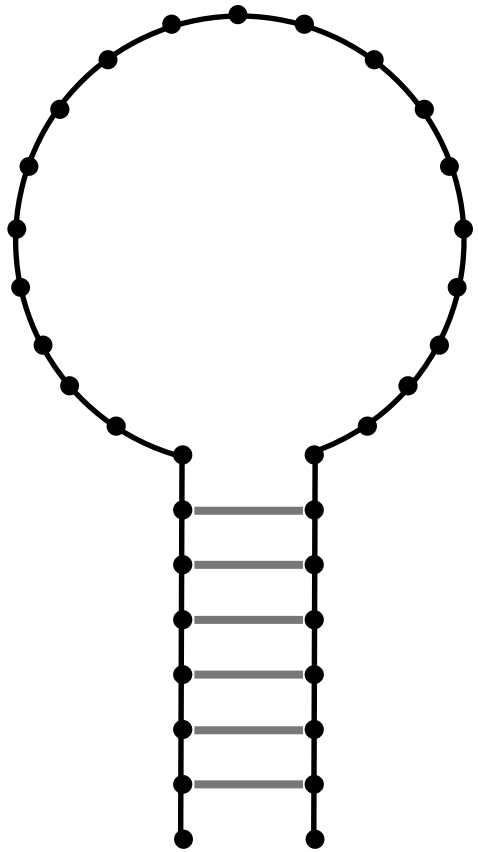
DNA

- just optimize base pairs (ask why later)

RNA

- base pairs
- stacking

  or
- count a contribution to loop

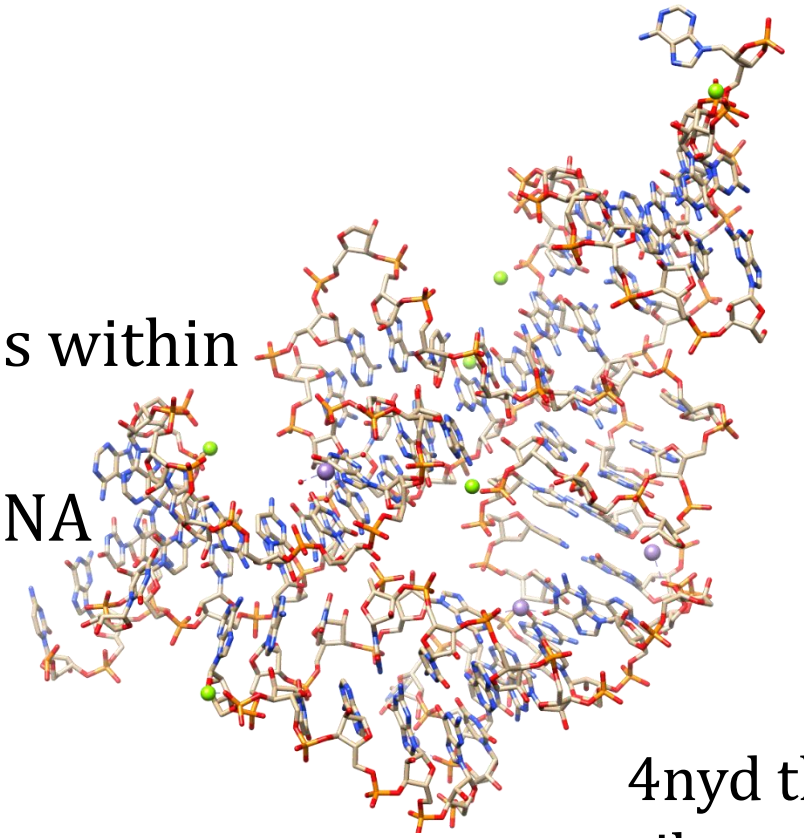


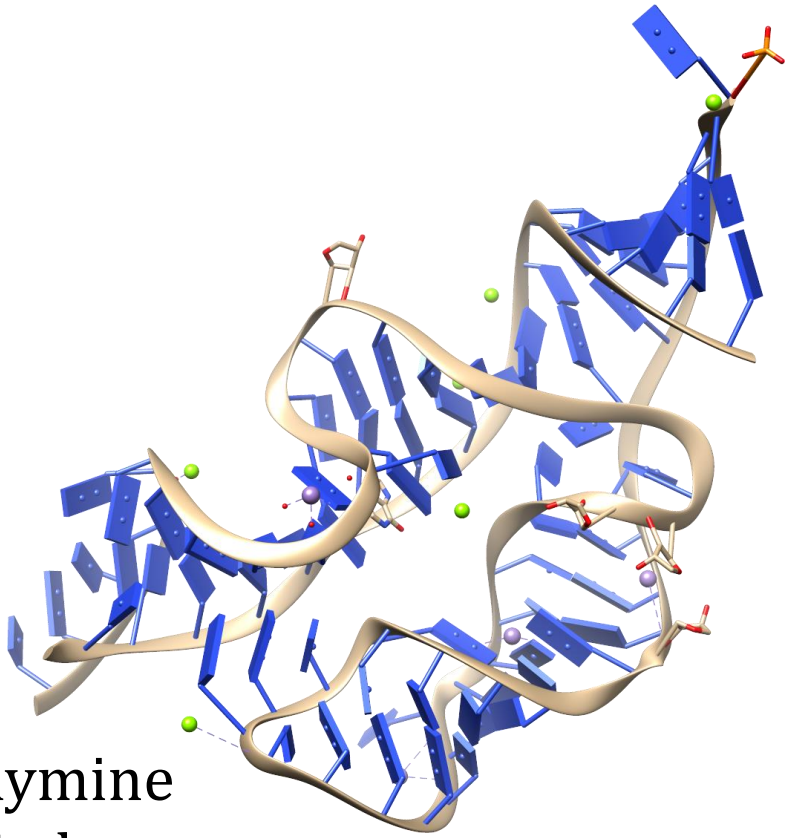
RNA Design

What does RNA do ?

- old view – information
- modern - information +
  - catalysis
  - binding / regulation

- likes to form double helices within one molecule
- much more flexible than DNA





4nyd thymine riboswitch

RNA Design

Similarities to protein design

- want to design compact structures from one strand (chain)
- size of problem ?
  - $4 \times 4 \times 4 \ldots = 4^n$ and a transfer RNA is about 75 bases (4^{75})

Special properties of RNA (contrast with proteins) – details coming

1. 2D description
2. simpler energy models
3. structure prediction

1. RNA 2D world

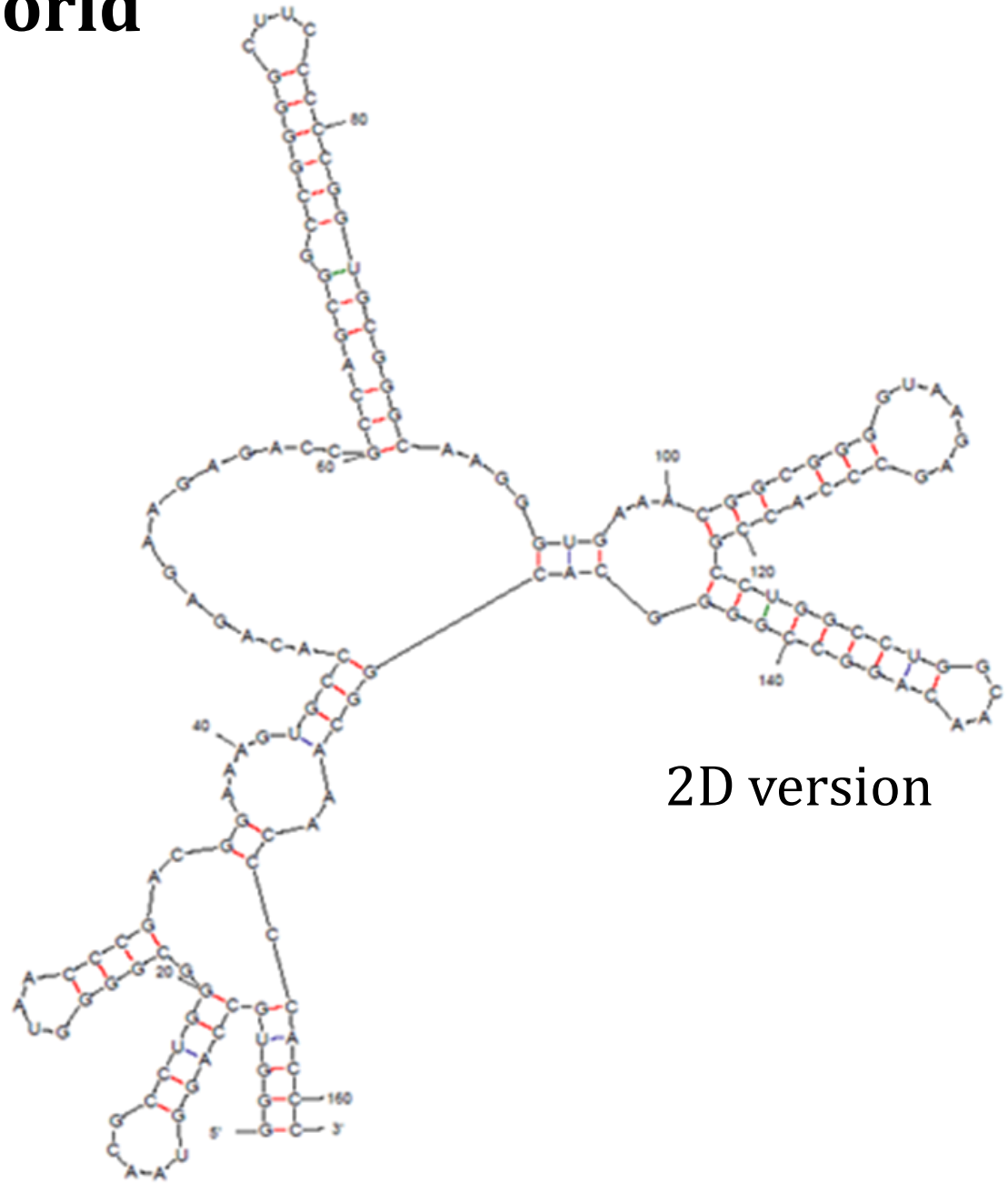
proteins

- 3D structures

RNA

- 2D literature

crystal structure
PDB acquisition code 1u9s



2D version

2D model consequences

proteins ?

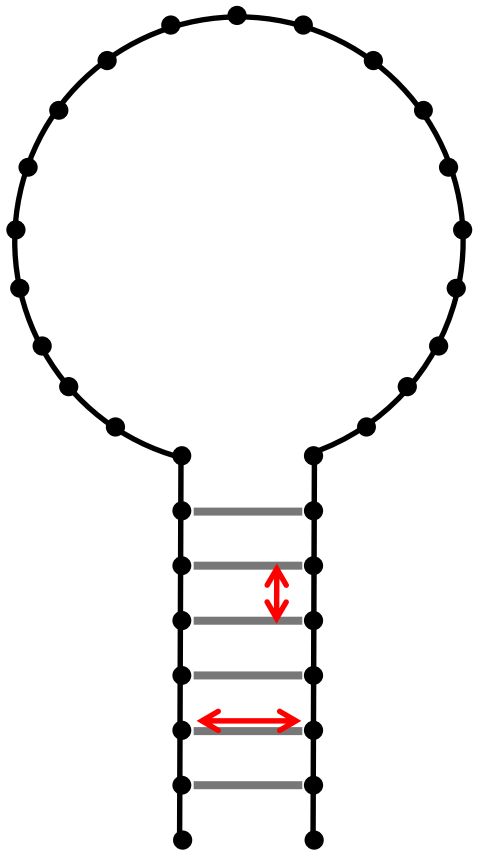
- an amino acid has n neighbours (n is some small number)

RNA

- neighbour across the base pair
- neighbour up and down in sequence

  or
- no neighbour (count loop contribution)

- for a given structure – number of neighbours is very small
- no sidechain geometry (ignored / averaged)



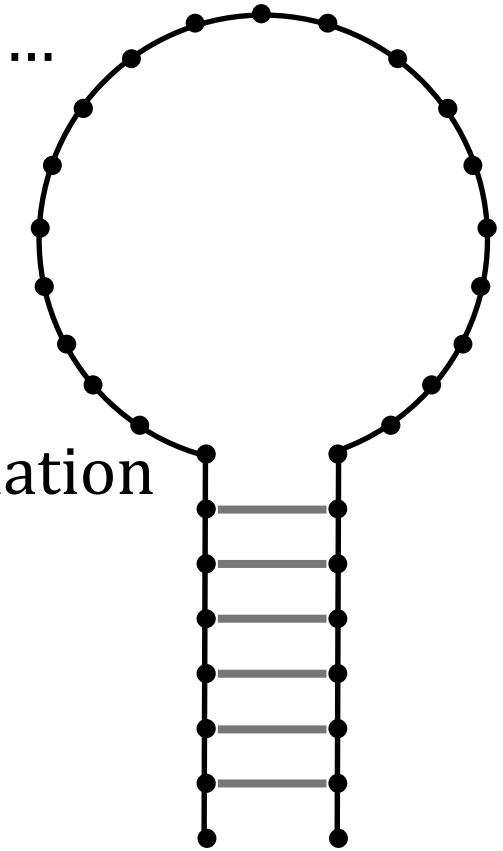
2. RNA – simple energy model

Proteins

- nearly always distance dependent - $\frac{q_i q_j}{4\pi\varepsilon r_{ij}}$, $4\varepsilon\left(\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6}\right)$...

RNA

- discrete – what are the bases in a particular interaction ?
- easier problem – do not have to worry about details of conformation



3. RNA structure prediction

Proteins

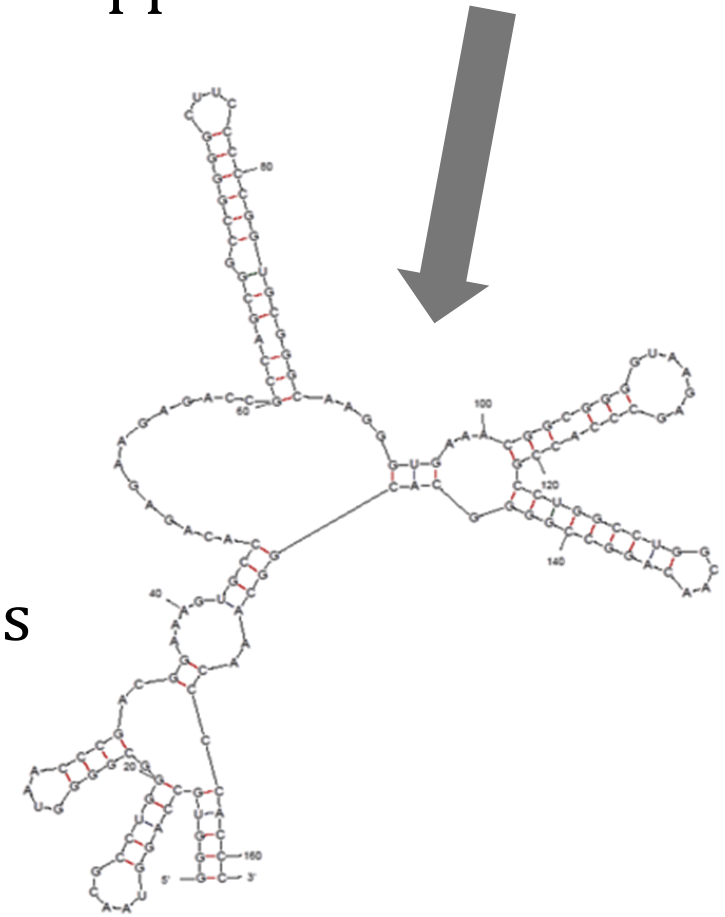
- cannot really reliably predict structure
- change an amino acid and have no idea what will happen

RNA

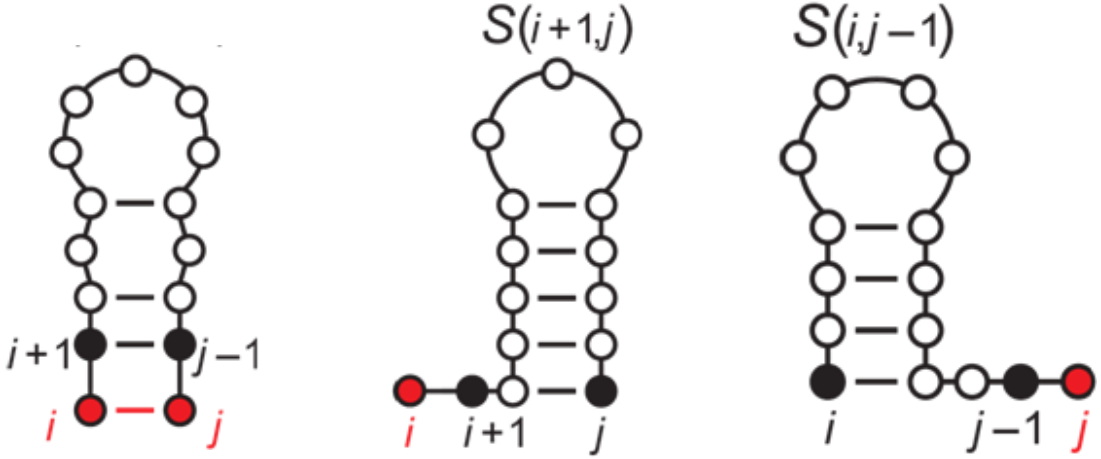
- different philosophy
- claim
  - you can predict 2D structure

- structure prediction is used in the design process (later)

ACGUACGG...



3. RNA structure prediction



- find optimal start of loops
- grow, allowing for gaps
- check for better scores by splitting loops

Result

- can find optimal 2D structure in $O(n^3)$ time

Is this true ? Can one really predict RNA structure ?

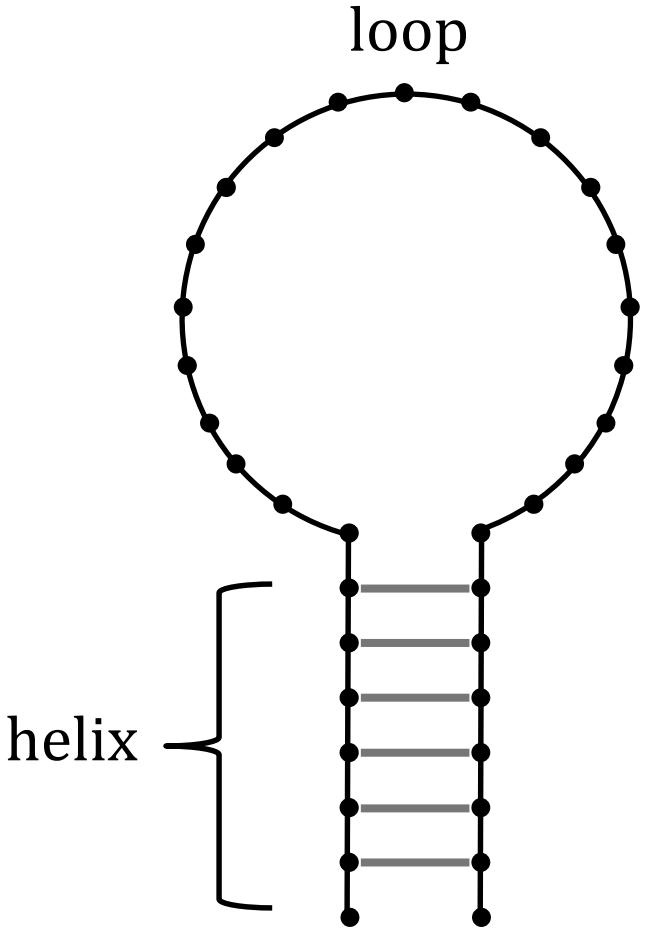
- as posed
  - yes – deterministic, optimal set of base pairs for a given score function
- physically
  - no – 20 – 25 % of predictions are very wrong
- does it matter ? – for today – no. Imagine we can predict structure

picture from Eddy, S.R. Nature Biotech 22, 909-911 (2004)

The energy model

- GC pairs score very well
- AU pairs score almost as well
- GU pairs score a bit
- neighbours in the chain get a score if they are in a helix
- details we ignore

Finally a design algorithm…



Towards sequence prediction

version 1, simple Monte Carlo

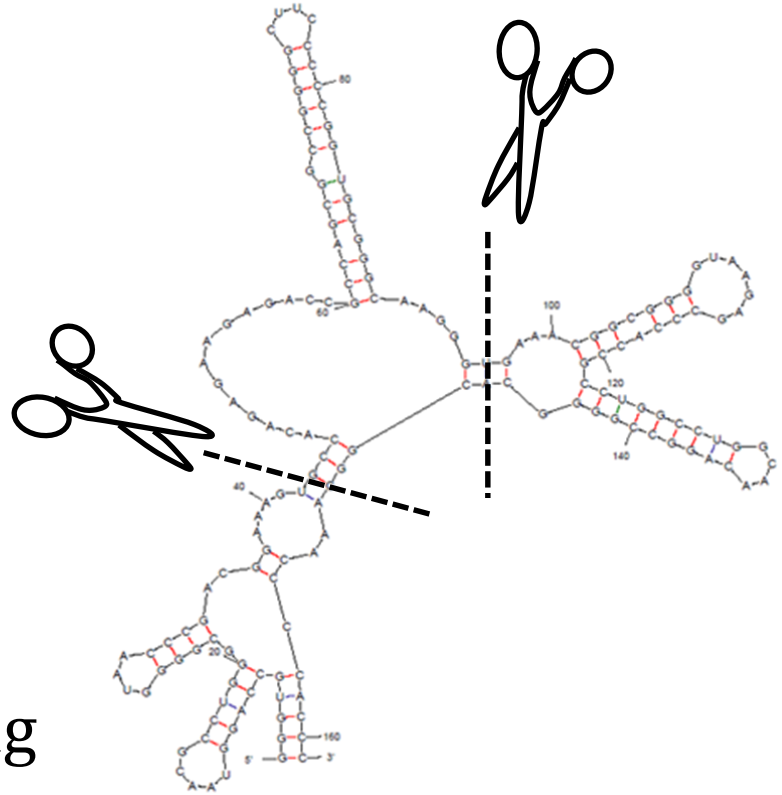
```
S = random sequence
while (not happy)
     change a base (S_trial)
     calculate ΔE
     if ΔE < 0
          accept S_trial
     else
          r = rand (0..1)
          if exp(ΔE/T) > r
               accept S_trial
```

why is this bad ?

Problems with simple Monte Carlo

1. size of search space
2. negative design

Search space



1. split molecule into pieces

Optimize separately and hope for no interactions

2. do not pick sites to change randomly

When generating S_{trial}, pick sites with wrong base pairing

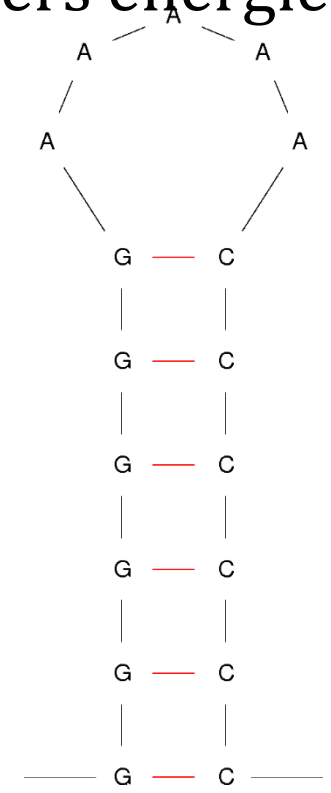
other words

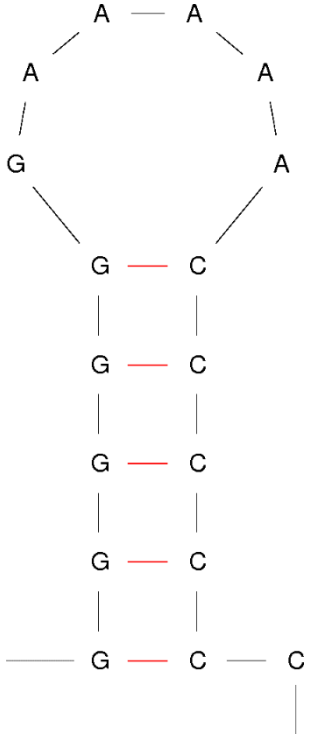
try not to break sites which seem happy

Negative design

- negative design = problem with alternative folds
- problem
- GC has 3 Hydrogen bonds, AU has 2 – what would be your solution ?

- same sequence – two answers energies almost the same





negative design – the problem

Same sequence – two equally good solutions

More generally

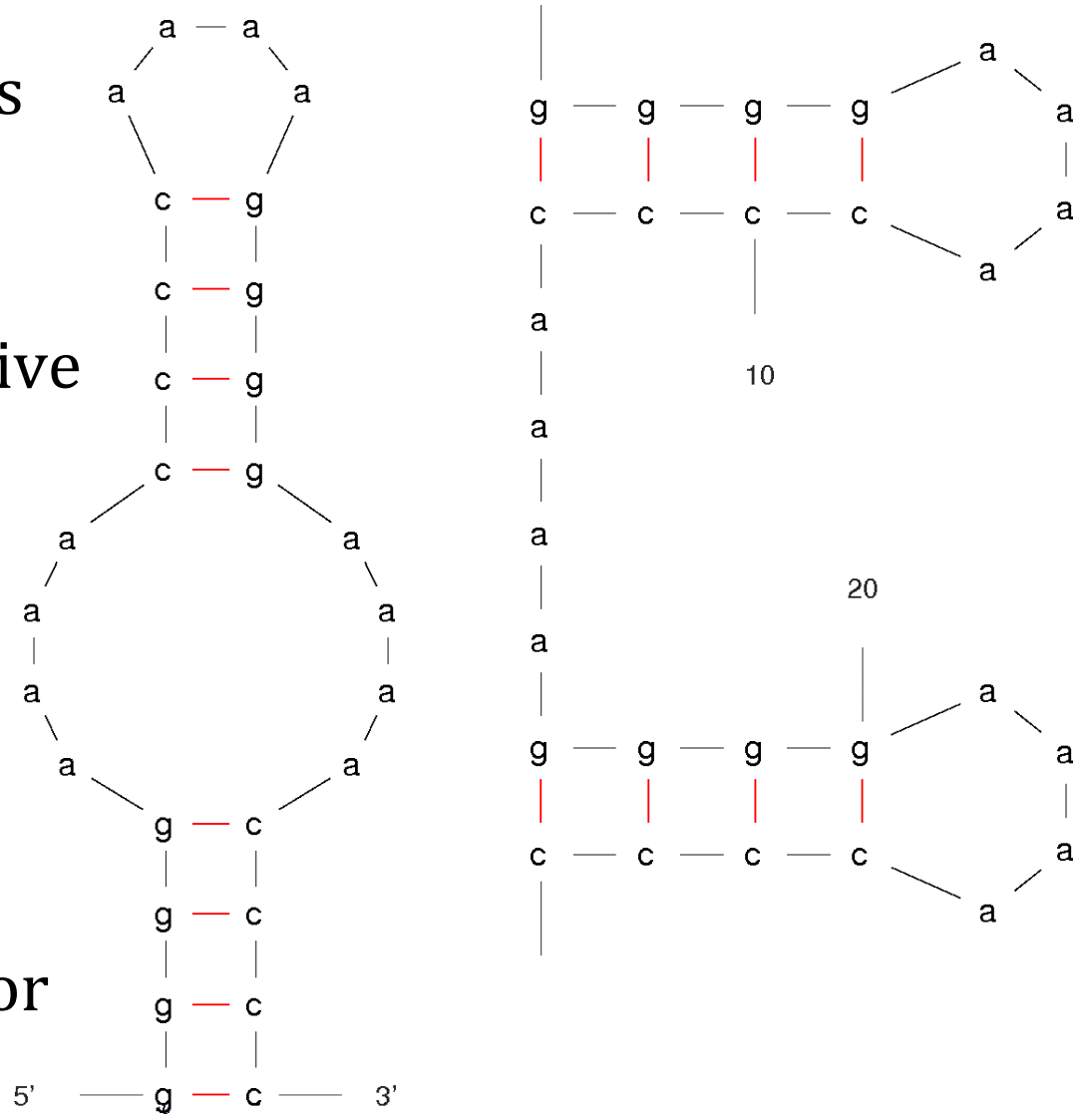
- naïve GC rich solutions will have alternative folds

What is negative design ?

- find a sequence which will not fold to wrong structure

New version of selection criterion – select for

- energy
- not folding wrongly



Final RNA design method

```
[ break into pieces ]
initial sequence
while (not happy)
     change residues
     calculate energy – reject ?
     calculate structure – accept / reject
```

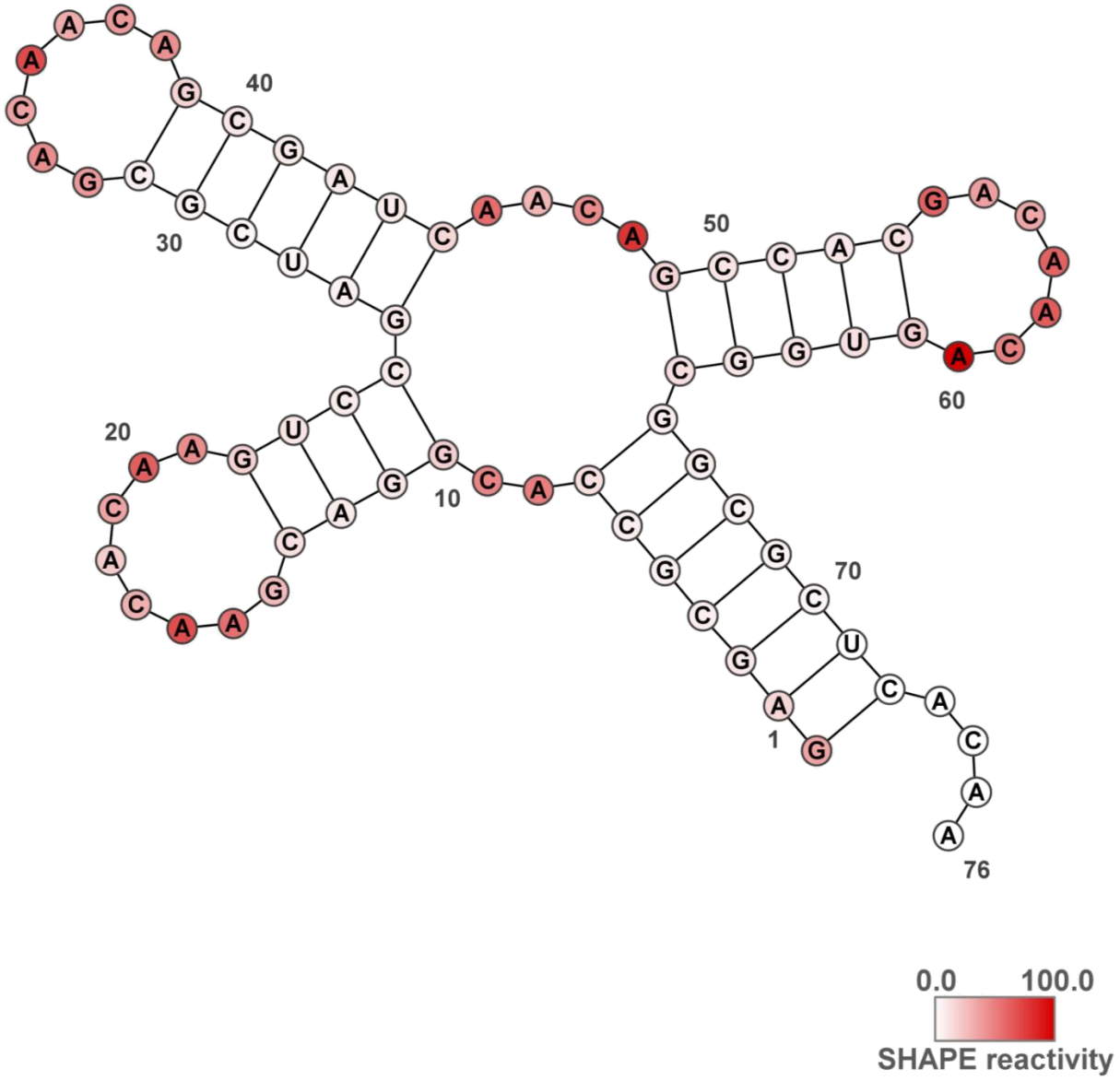
simple energy model

$O(n^3)$method mentioned earlier

Does it work ? – self indulgence

a designed sequence

- red means not in a base pair
- base pairs a mixture of GC and AU
- not a simple looking sequence



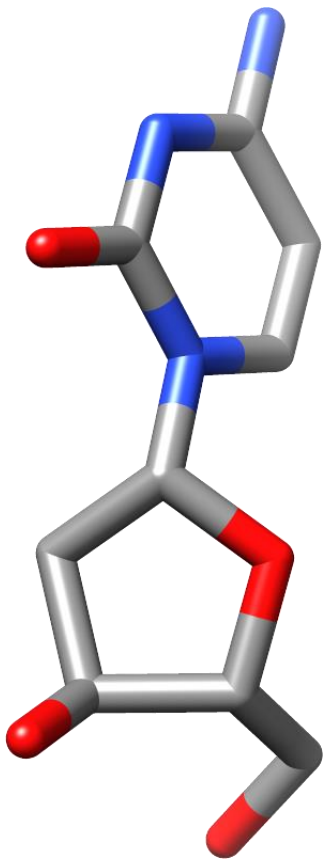
Enough RNA

RNA vs DNA

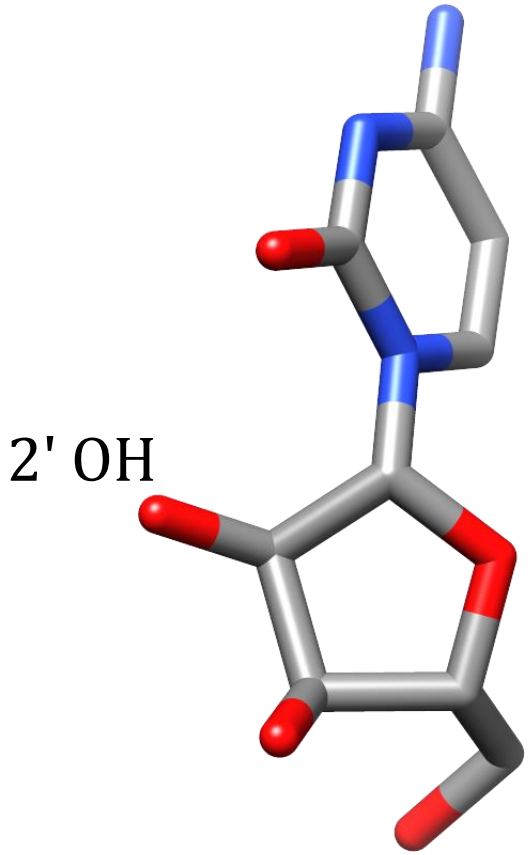
Chemical difference is small

DNA

- much less flexible
- nearly always helical



DNA (C)



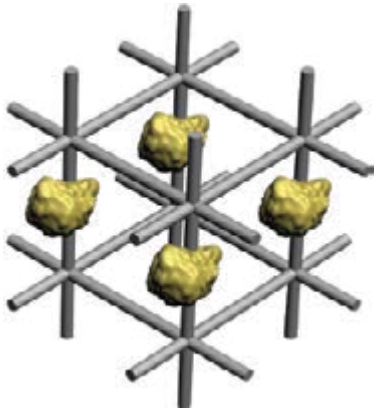
RNA (C)

DNA and templated design

Longer term aim – design long relatively simple shapes build scaffolds, boxes, ..

DNA
building tile

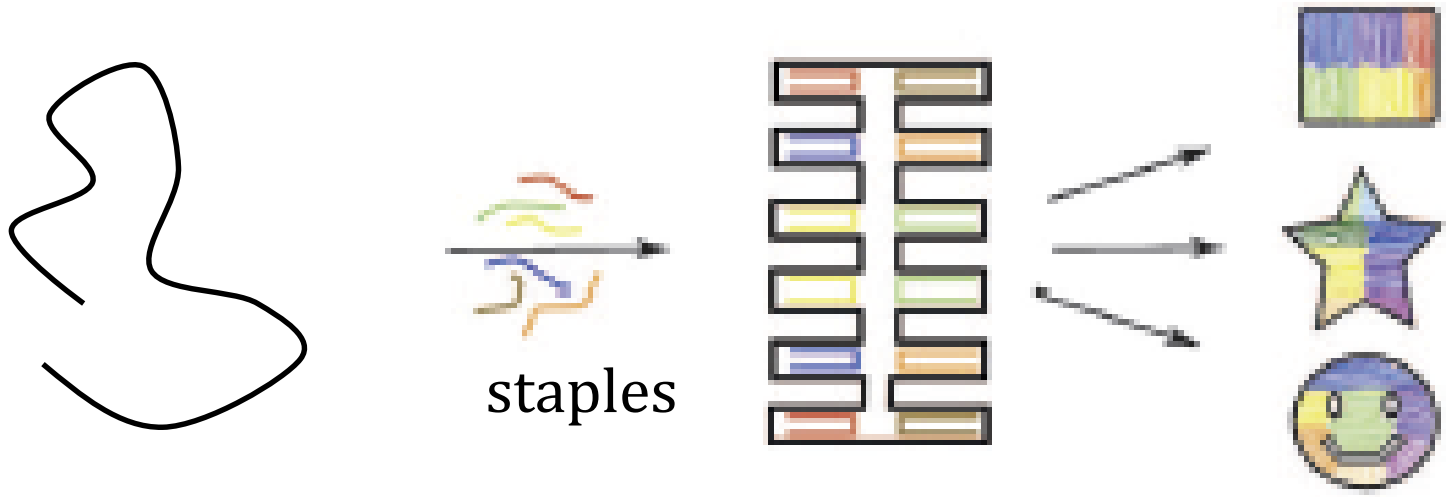
protein



Pinheiro, AV, Han, D, Shih, WM, Yan, H., Nature Nano 6, 763-782

scaffold philosophy

10^3 bases – natural DNA



scaffold DNA

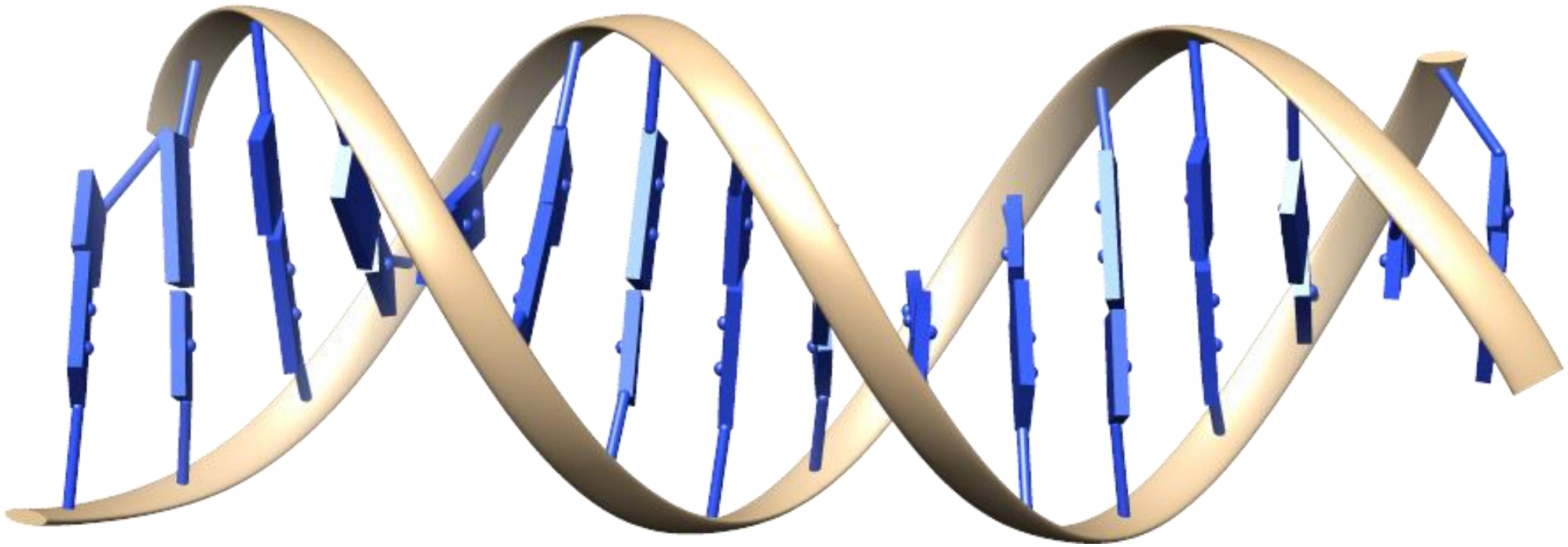
assemble by complementarity

details of first DNA origami

Somoza, A., Angew. Chem. Int. Ed. 48, 2-5, 2009

DNA origami

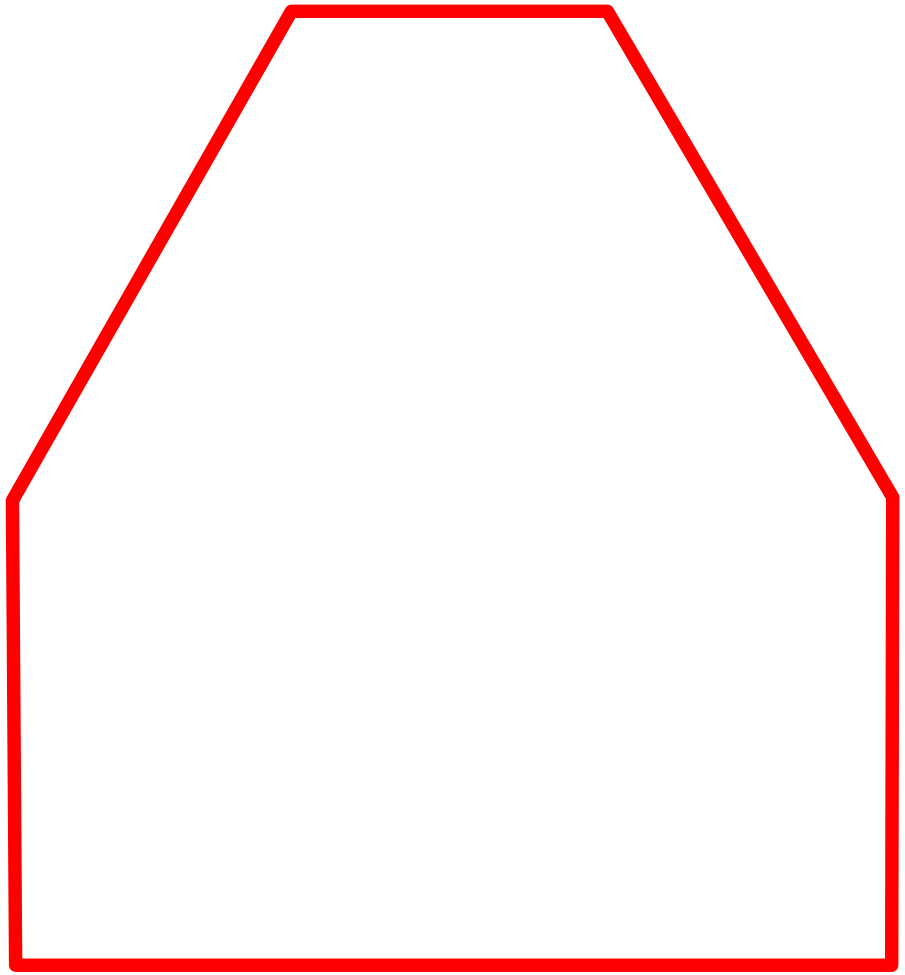
Remember DNA is most stable as a double helix



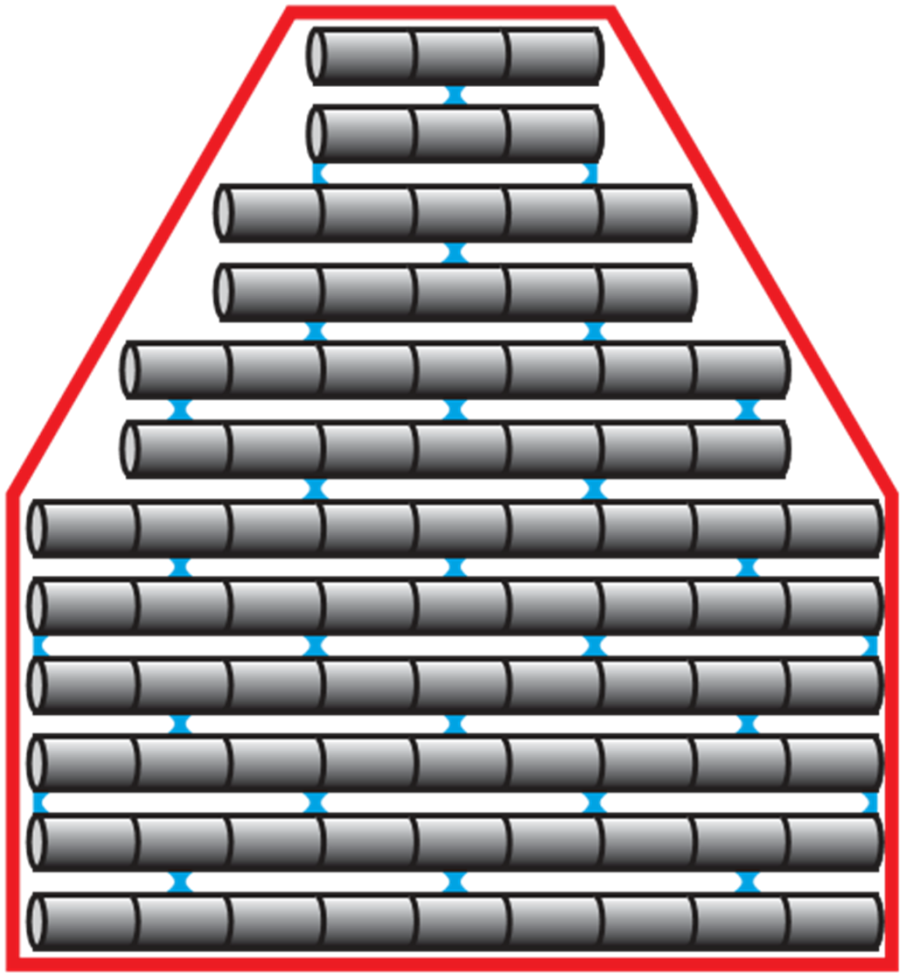
20 Å

36 Å

one turn, 10 ⅔ base pairs



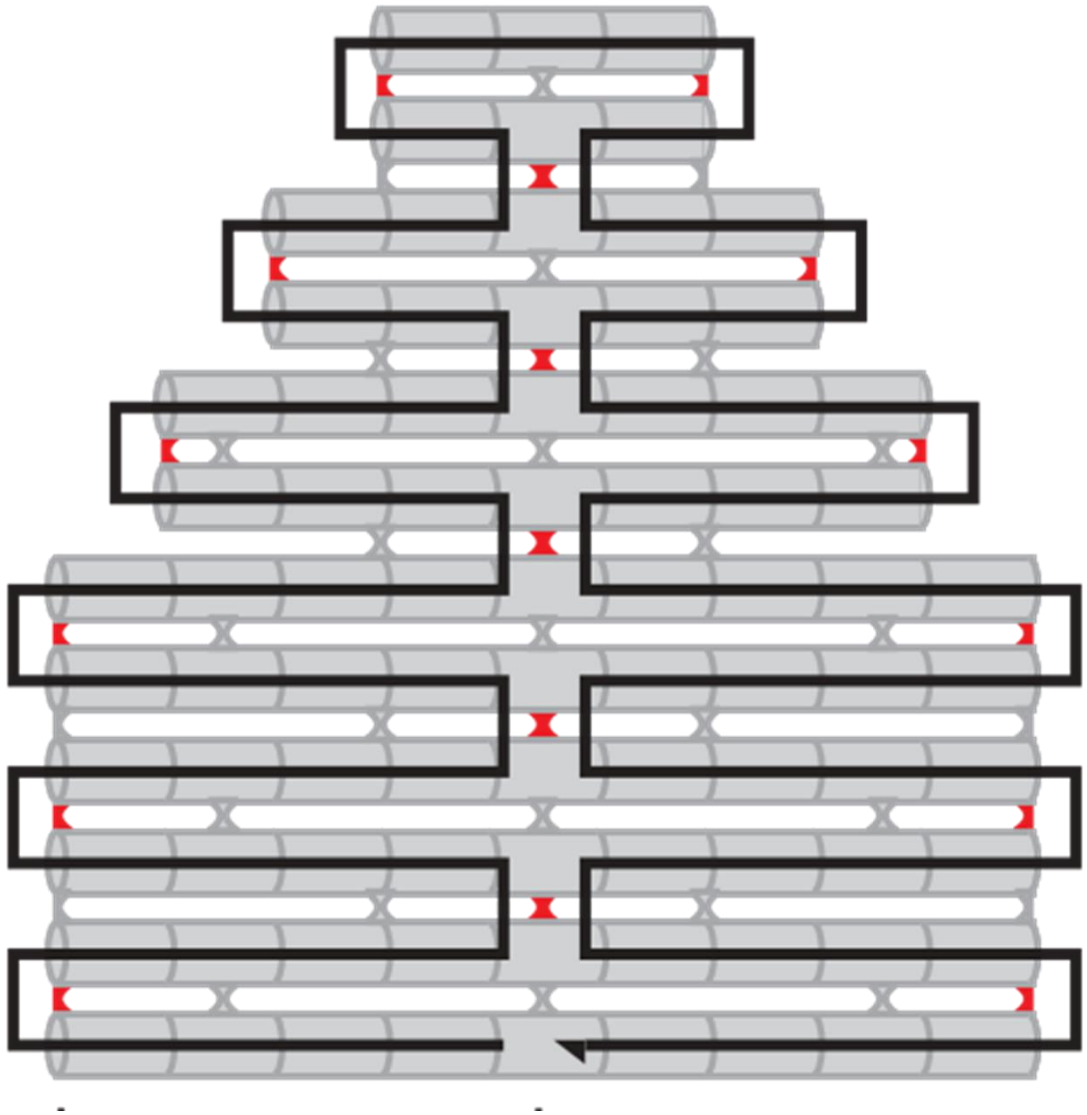
decide on shape

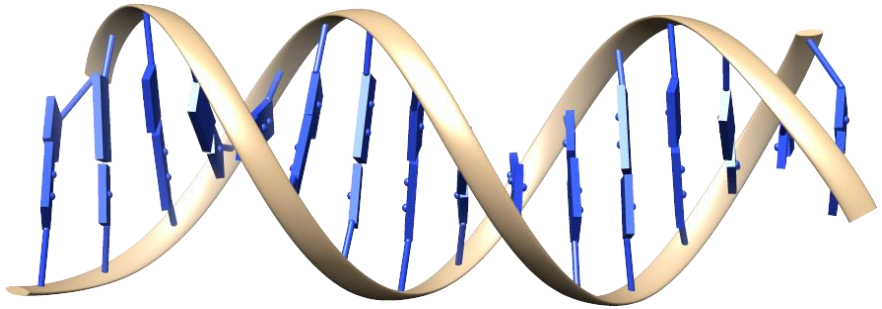


fill with cylinders
20 Å thick
length $\times \frac{10\ ⅔}{36}$ bases

Rothemund, PWK, Nature, 440, 297-302, 2006

One long strand runs along structure





Every ½ turn brings other chain into position for crossing over…

Rothemund, PWK, Nature, 440, 297-302, 2006

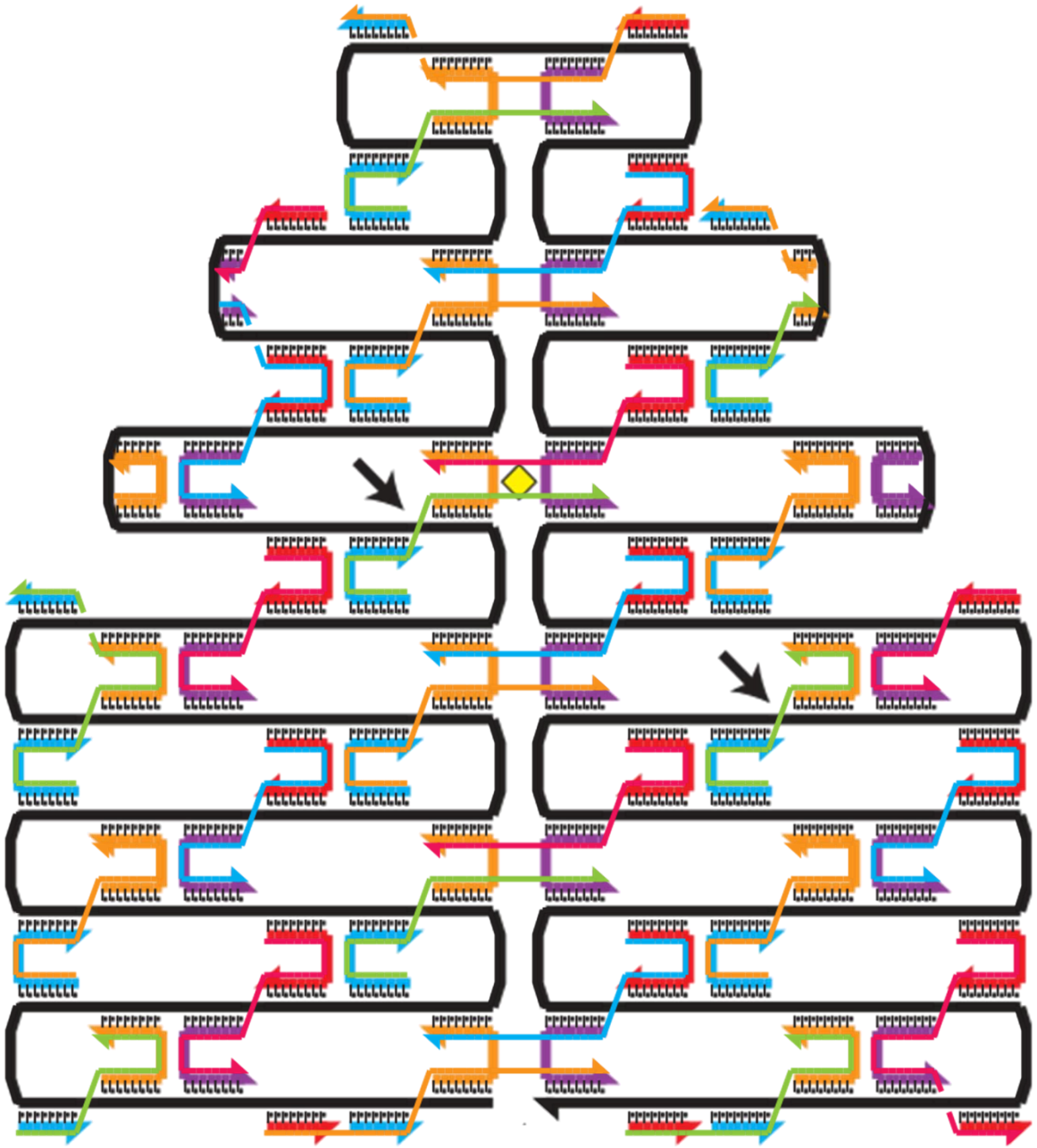
place joining strands (staples)

then join the staples into longer pieces..

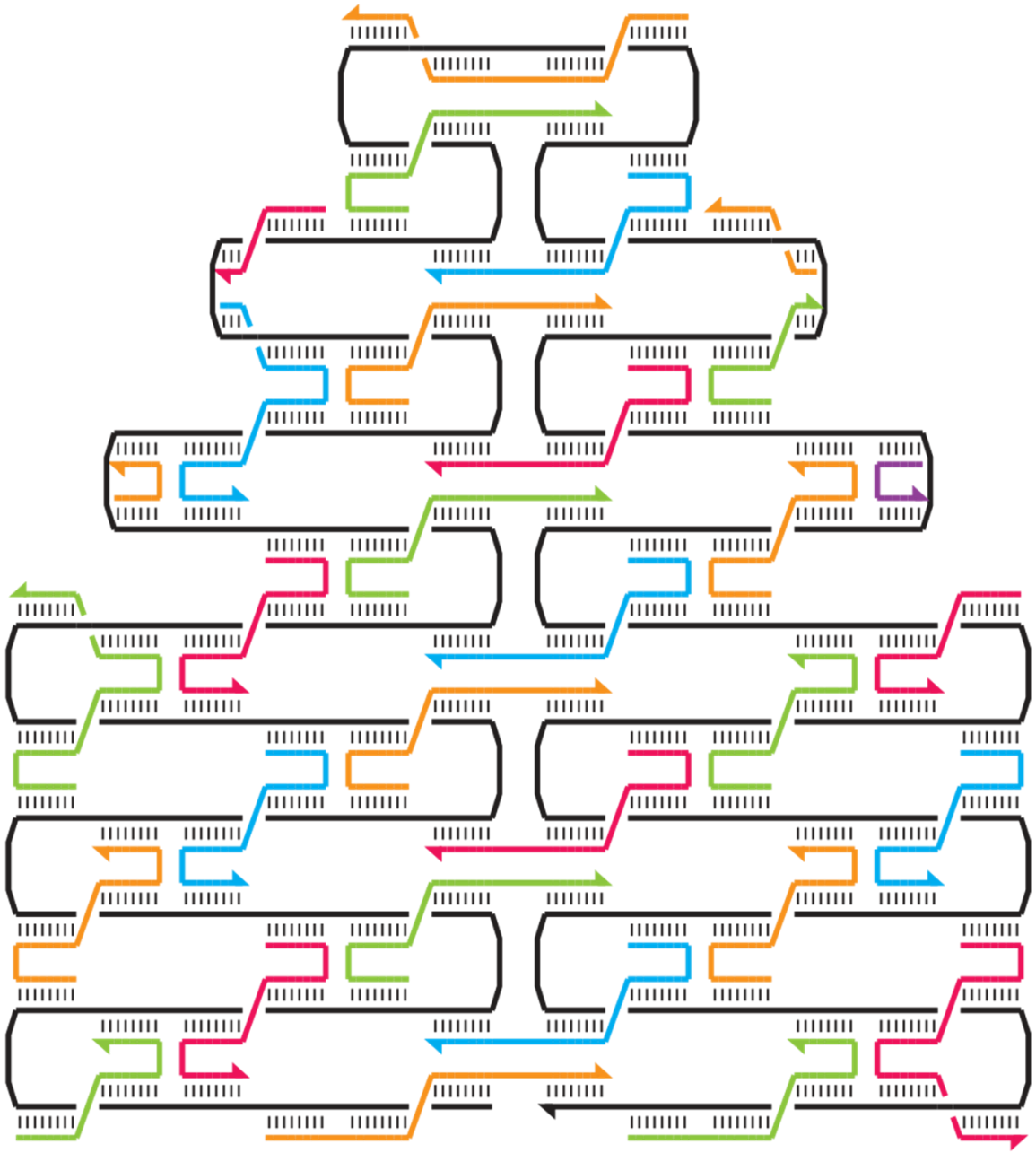
detail

every base is paired

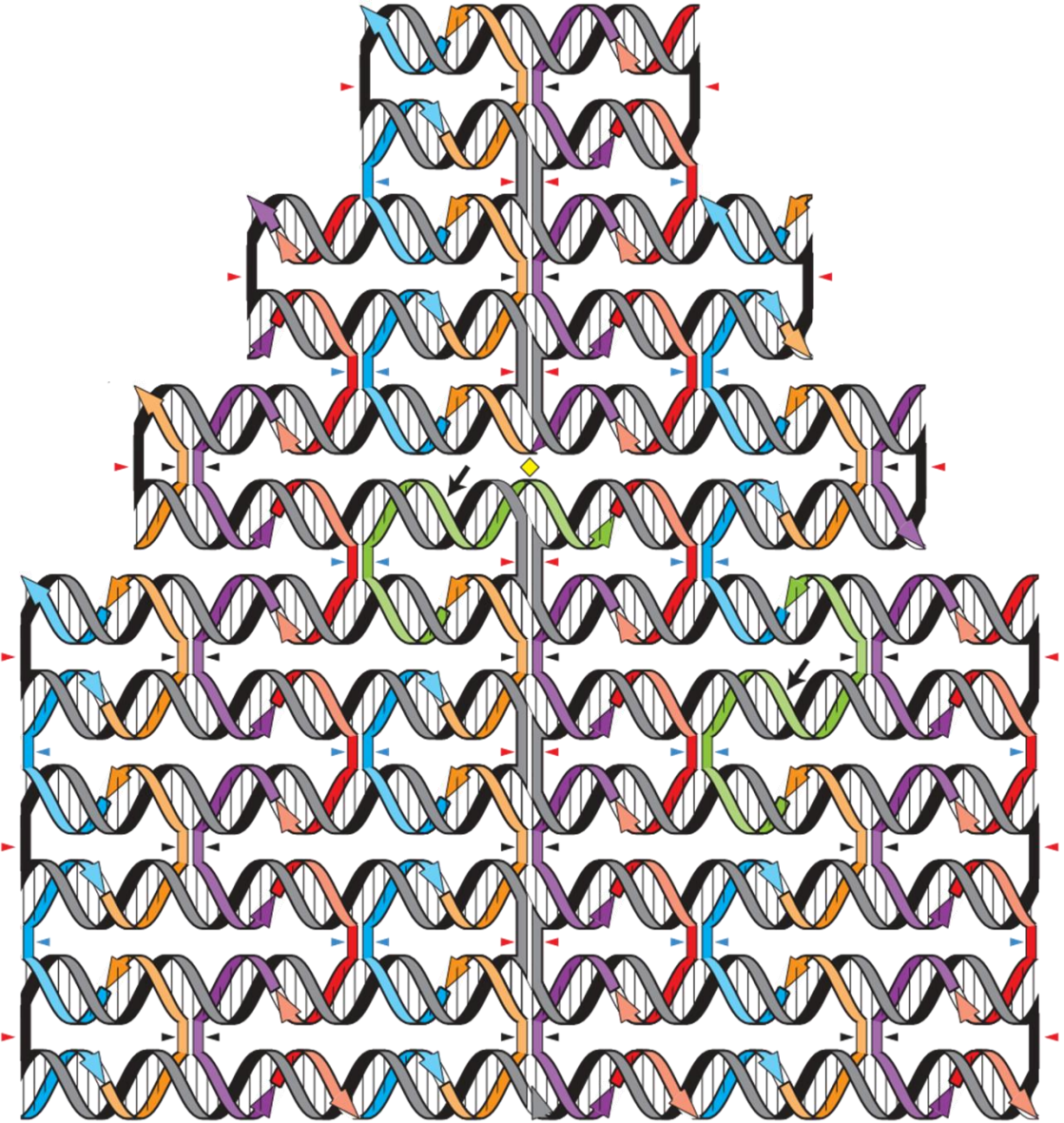
Next look at staples and join them

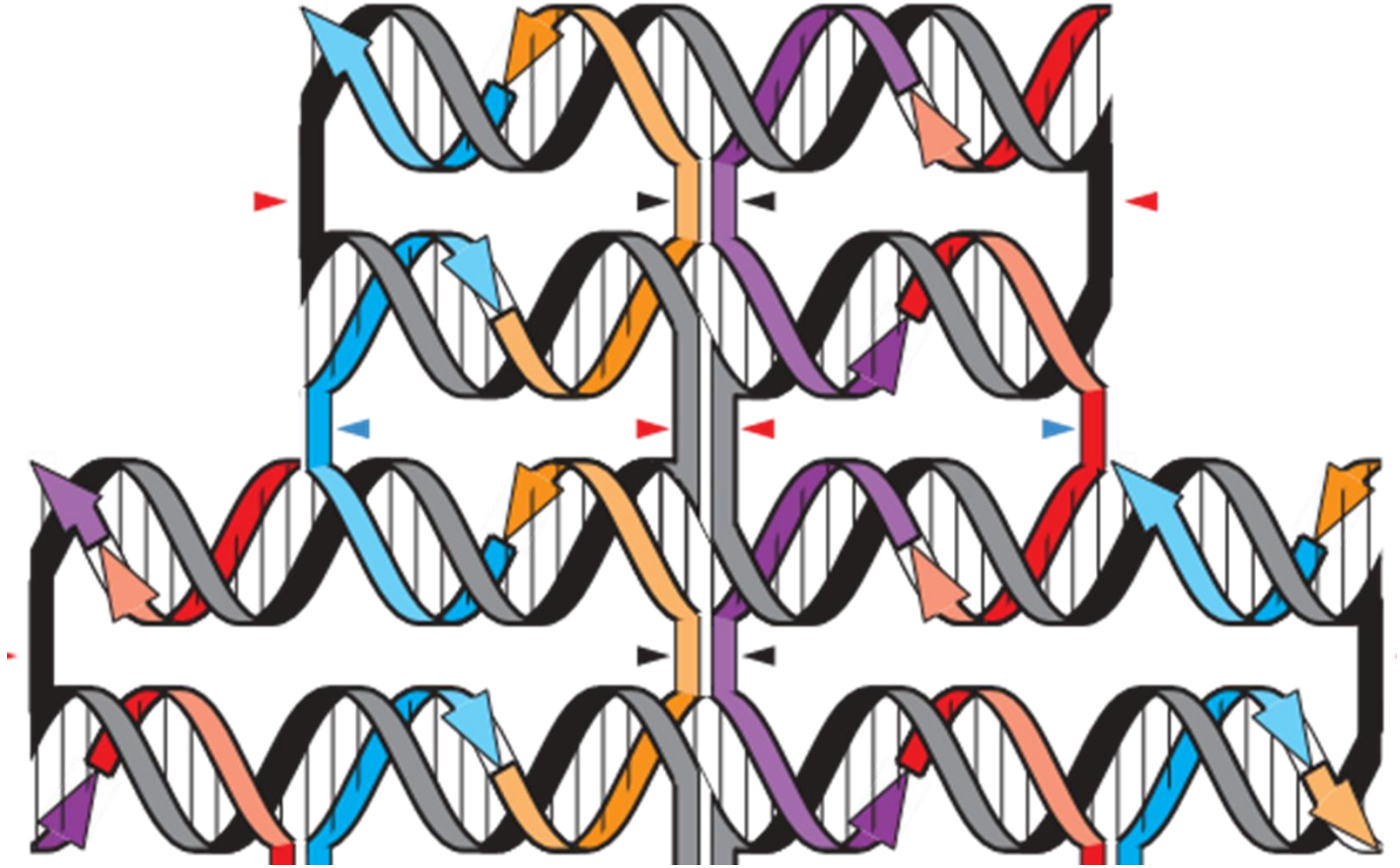


Rothemund, PWK, Nature, 440, 297-302, 2006



Rothemund, PWK, Nature, 440, 297-302, 2006



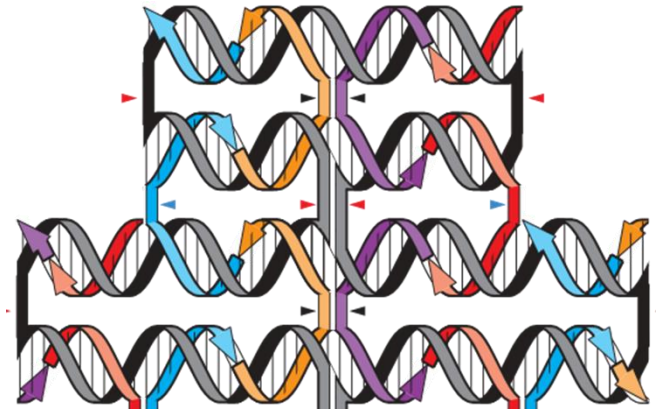


basically a long double helix
one long strand
lots of staple/joining strands

Rothemund, PWK, Nature, 440, 297-302, 2006

details of DNA origami

- program makes list of staple sequences
- units ?
  - helices are in units of ½ turns



Self assembling

- throw long strand + joiners into a bucket and let it reassemble

Rothemund, PWK, Nature, 440, 297-302, 2006

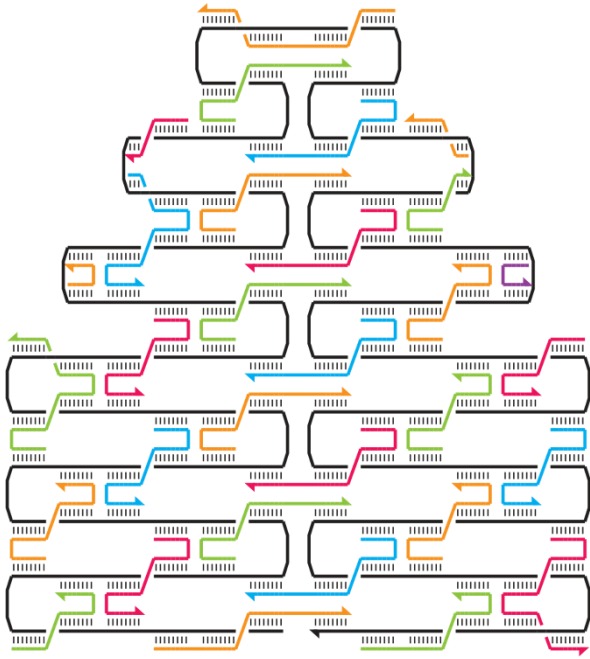
where are we ?

In this style of design

- long DNA strand is
  - taken from nature (phage)
  - not really designed

- short staple strands
  - are designed
  - staple / heften / hold together the long strand in some shape

negative design



Where is the "negative design" ?

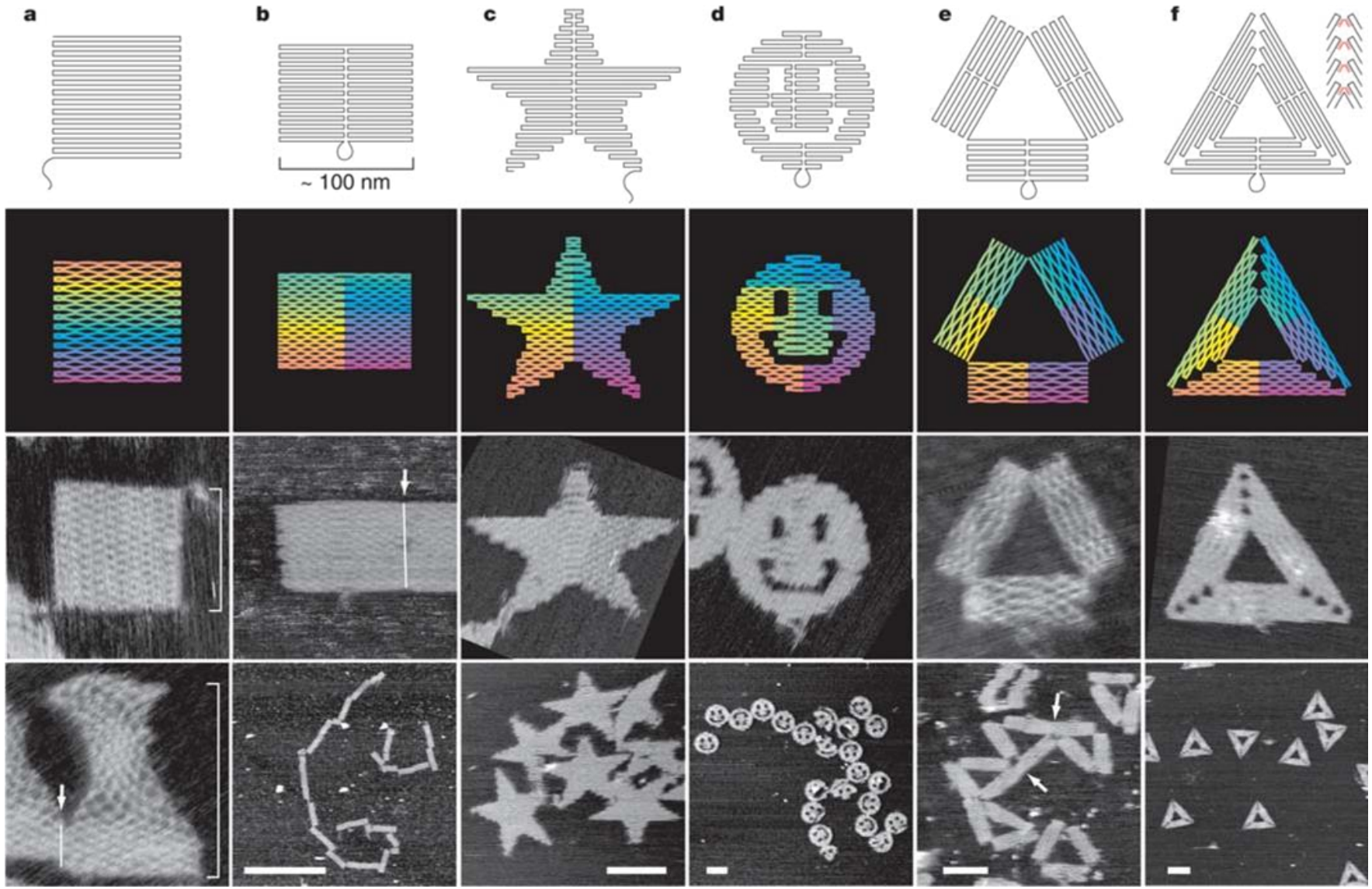
- you have a large natural piece of DNA – no repetitive elements
- staples fit to a specific part of long strand – not to other parts

Is this true ?

- true enough (procedure works - next slide)
- what really happens – building structures takes hours not seconds

  - joining staples match best to target regions – weakly elsewhere
  - gradually cooling a system lets staples usually find best match

Rothemund, PWK, Nature, 440, 297-302, 2006



designed shape

designed chain coloured

microscopy

microscopy

Rothemund, PWK, Nature, 440, 297-302, 2006

Summarise some properties

	DNA	RNA
	nano-scale	molecular structures
catalytic activity ligand binding	rare	common
	template design	*de novo*

	DNA	RNA
	double stranded	single / sometimes double
	GC, AT	GC, AU (+more)
	stable	not stable very sensitive to RNAse can be modified 2'-O methylation
ΔG energy per base per stack, kJ Mol^{-1}	−1.4	−3.6 to −8.5
synthesis	cheap	not so cheap up to 100 bases

Summary and stop

Remember differences

- protein vs nucleotide
- RNA versus DNA

- philosophy of energy functions

- differences scaffolded and *de novo* design

- could you design absolutely everything using a scaffolded method ?