Nucleotide design

RNA and DNA – very different application areas

- DNA
  - big structures (10^3 bases) – make a chain fold up
  - stable
  - long double helical shapes
  - scaffolding
- RNA
  - less stable
  - smaller structures (10^2 bases)
  - very specific structures / shapes
  - therapeutics / catalysts

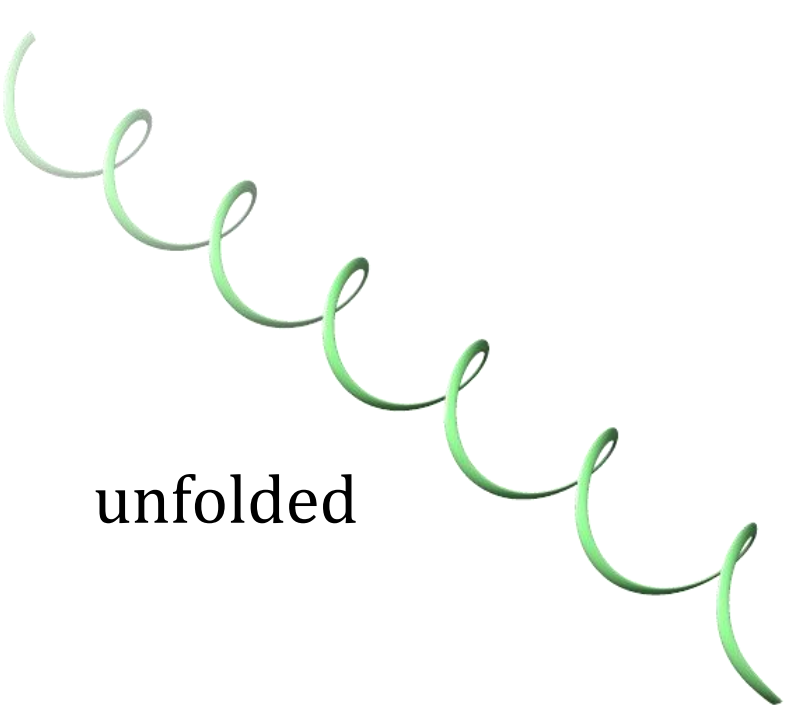
Feasibility

Nucleotide synthesis

- starting points / small quantities are cheap
- < 10^2 bases easy, 10^3 bases feasible

start with RNA

Naïve version RNA design



unfolded

ΔG

folded

In nearest neighbour model

ΔG is parameterised for exactly this

picture will become more complicated later

how to solve ?

The design problem (at this point)

- find a sequence to get best ΔG on target structure

- How big is search space ? $4 \times 4 \times \cdots = 4^{n_{bases}}$

version 1 simple Monte Carlo

score function is a ΔG, so comparing two values give $\Delta\Delta G$

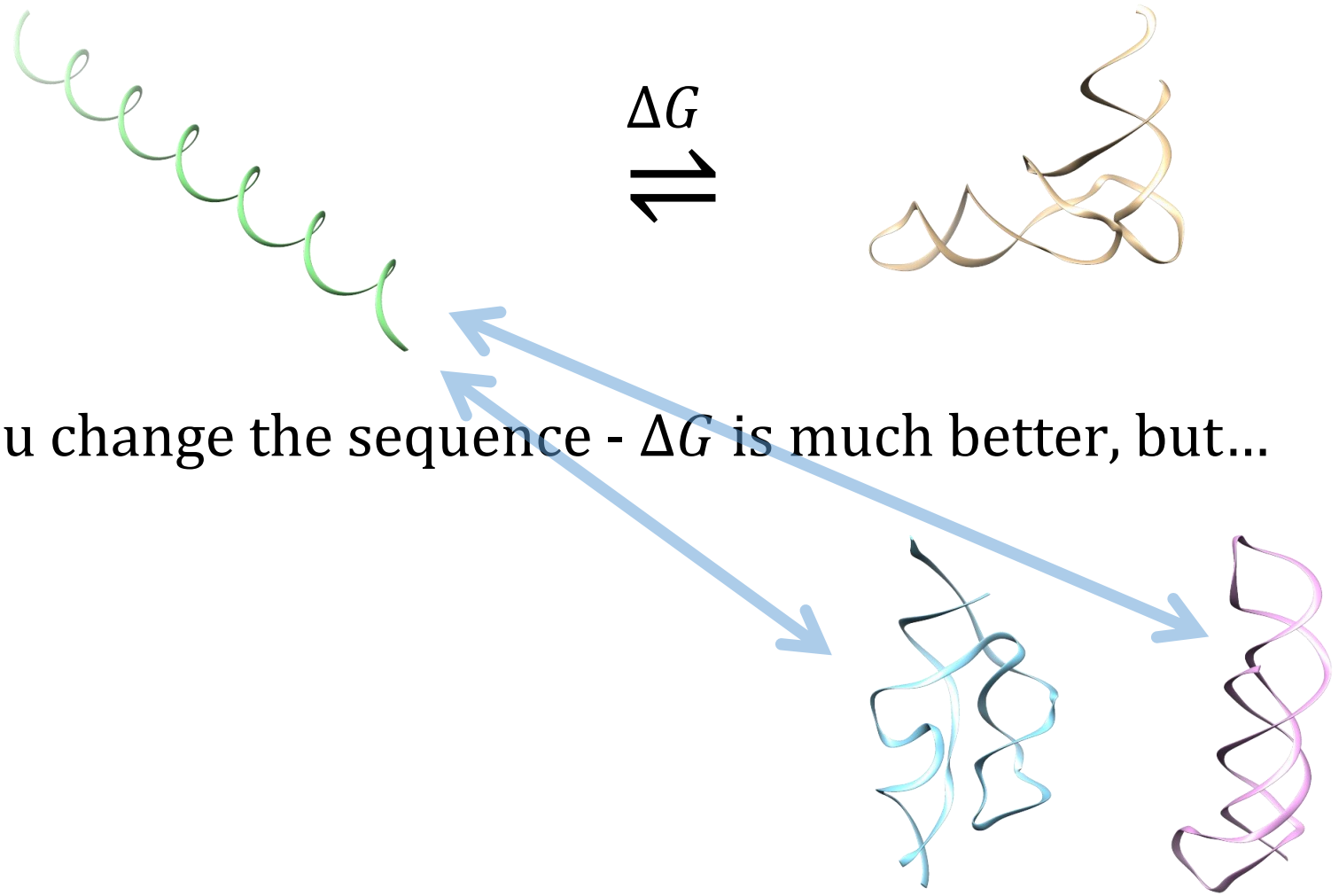
S = random sequence
while (not happy)
    change a base (S_{trial})
    calculate ΔG
    if $\Delta\Delta G < 0$
        accept S_{trial}
    else
        r = rand (0..1)
        if $\exp\left(\frac{\Delta\Delta G}{T}\right) > r$
            accept S_{trial}

Problems with simple version

1. search space is very big ($4^{n_{base}}$)
- ignore for today

2. negative design – wrong folding
- …

negative design



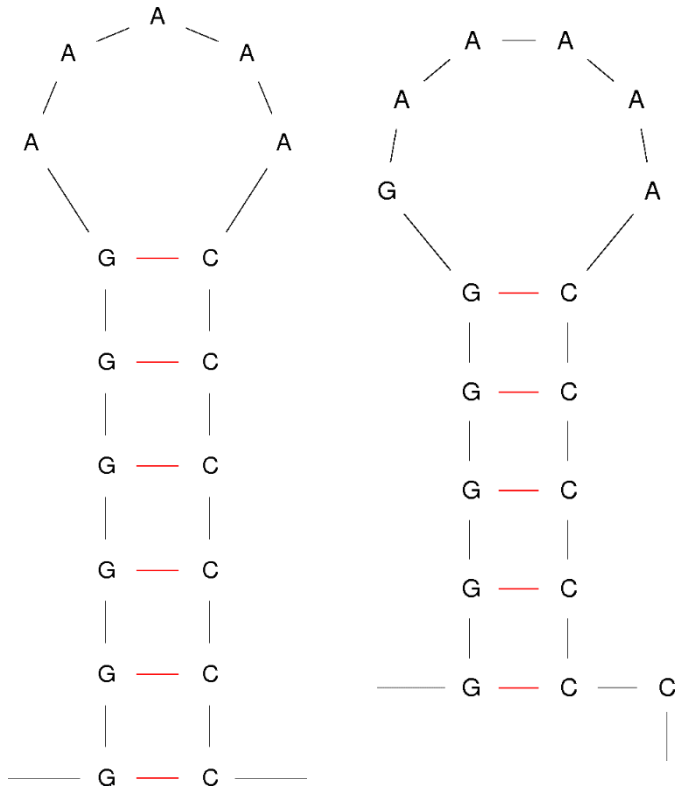
you change the sequence - ΔG is much better, but…

new sequence finds a different structure

How likely ? very ..

Negative design = problem with alternative folds

Same sequence
two answers energies almost the same



- same sequence – two good structures

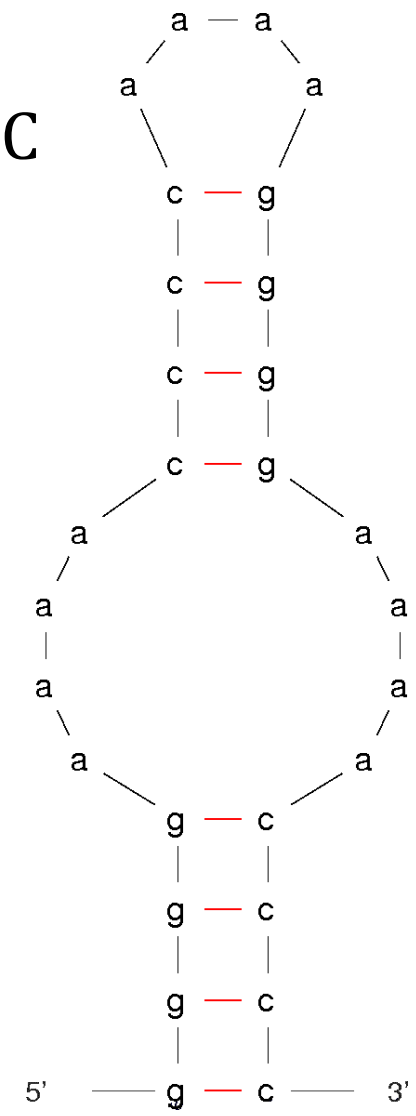
more generally

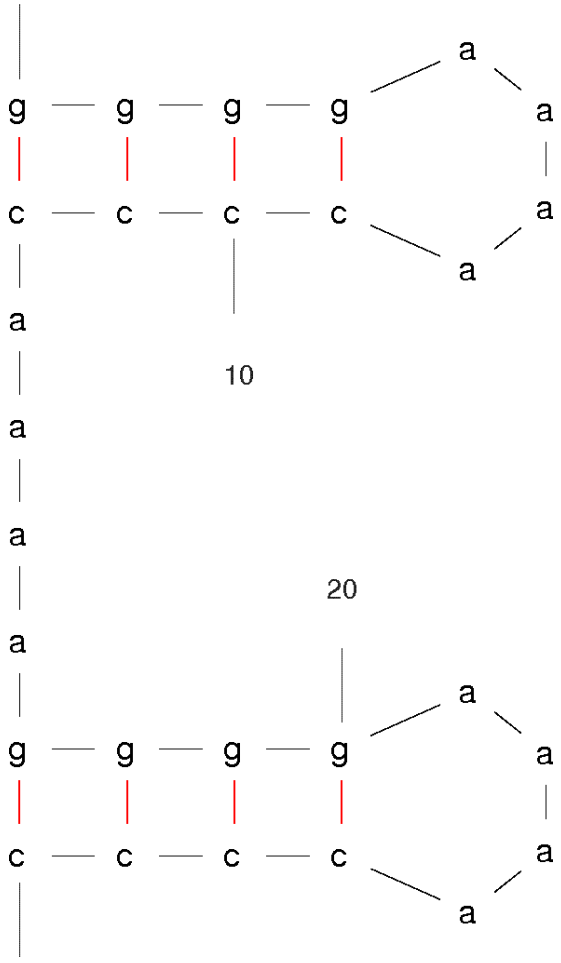
- ΔG is better with more GC pairs

if I have lots of GC pairs

- lots of alternative structures

- can we fix ?





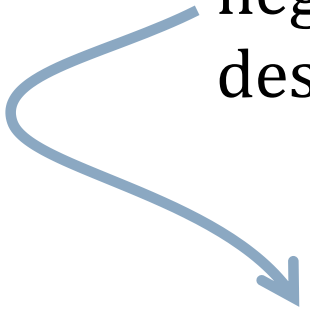
Fix problem with alternative structures

We optimised for $\Delta G_{folded-unfolded}$ - wrong value

```
while (not happy)
      change residues
      calculate energy – reject ?
```

- better version

```
while (not happy)
      change residues
      calculate energy – reject ?
      calculate structure – accept / reject
```



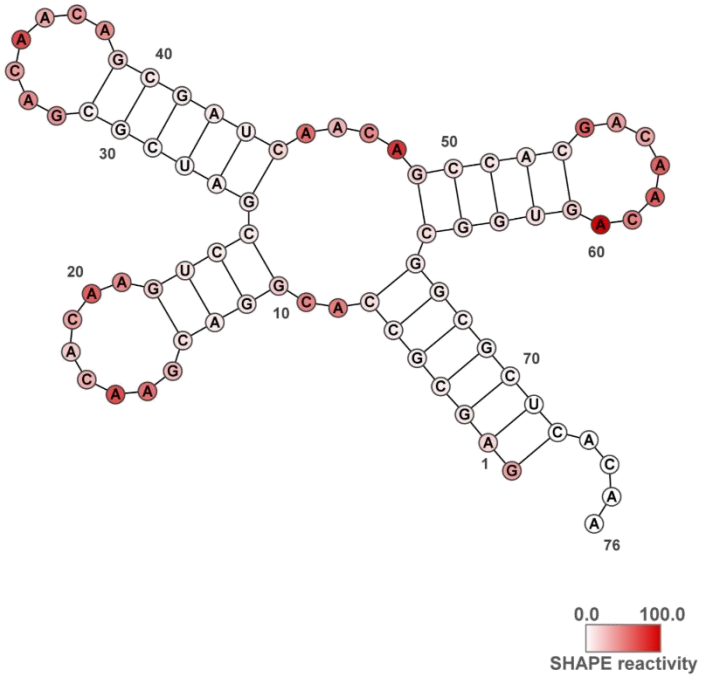
negative design

At every step – do a structure prediction calculation

- check if the target is still the optimal structure

Does it work ?

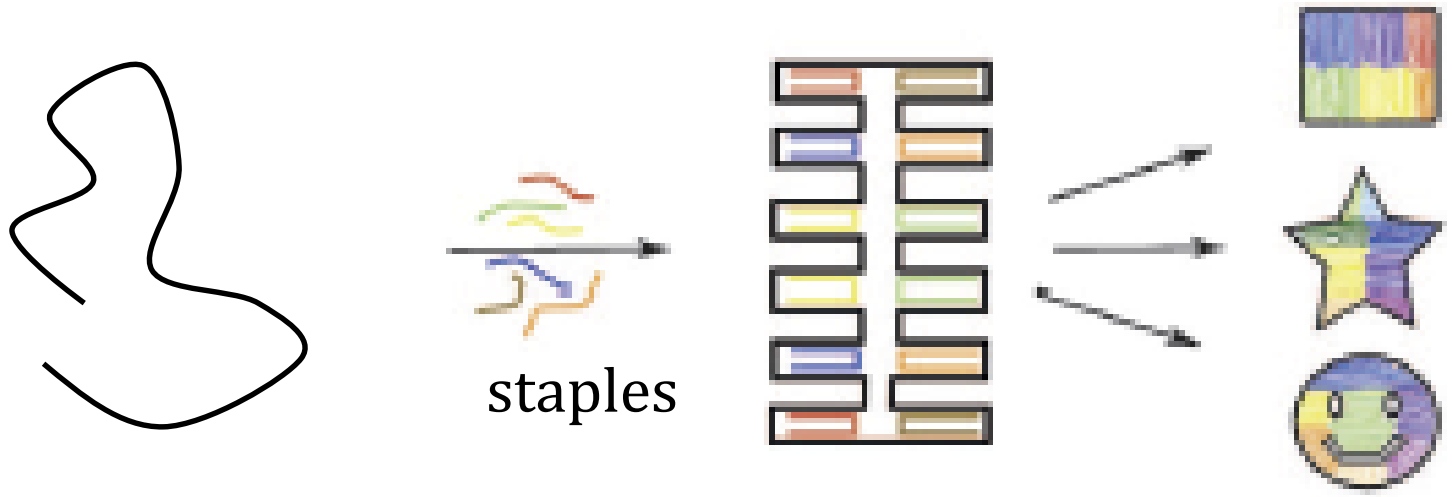
- red means not in a base pair
- base pairs a mixture of GC and AU
- not a simple looking sequence



What is broken ?

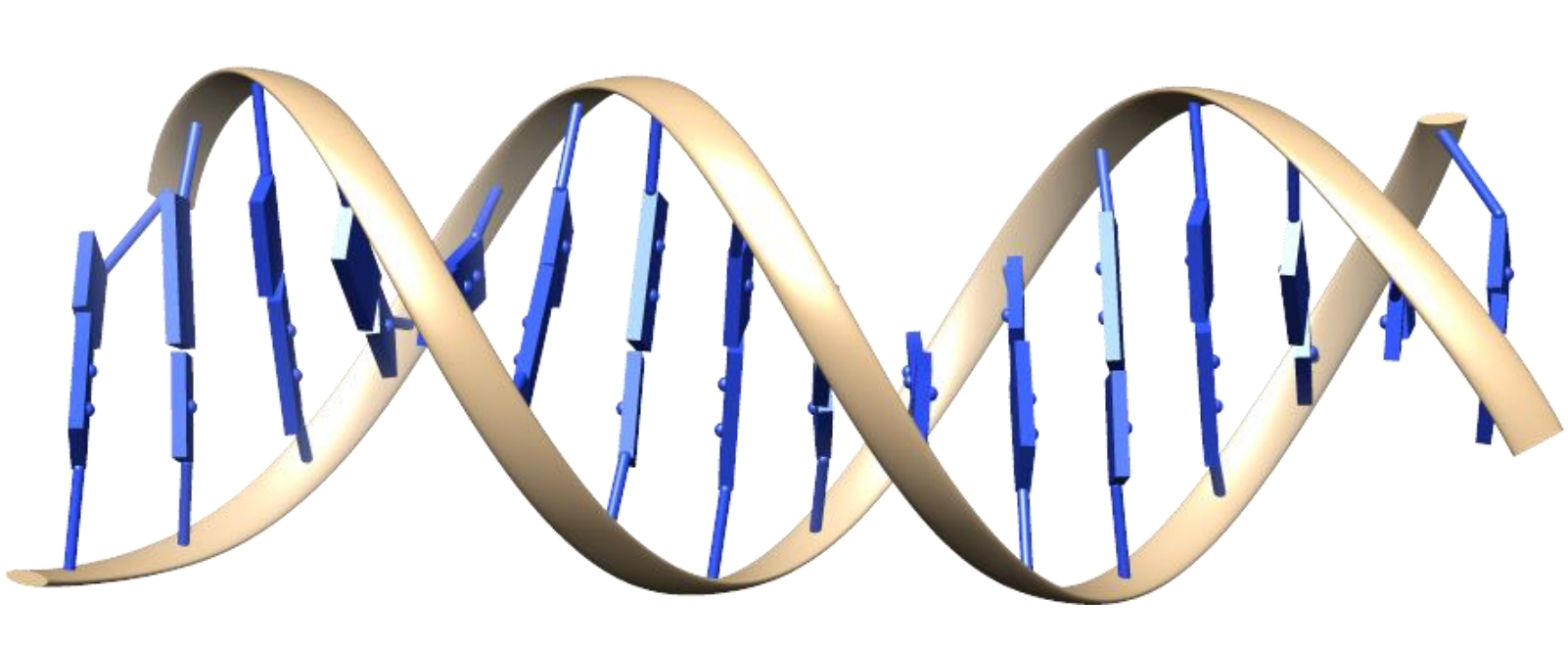
- requires structure prediction
- cannot do pseudoknots

DNA design – very different



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

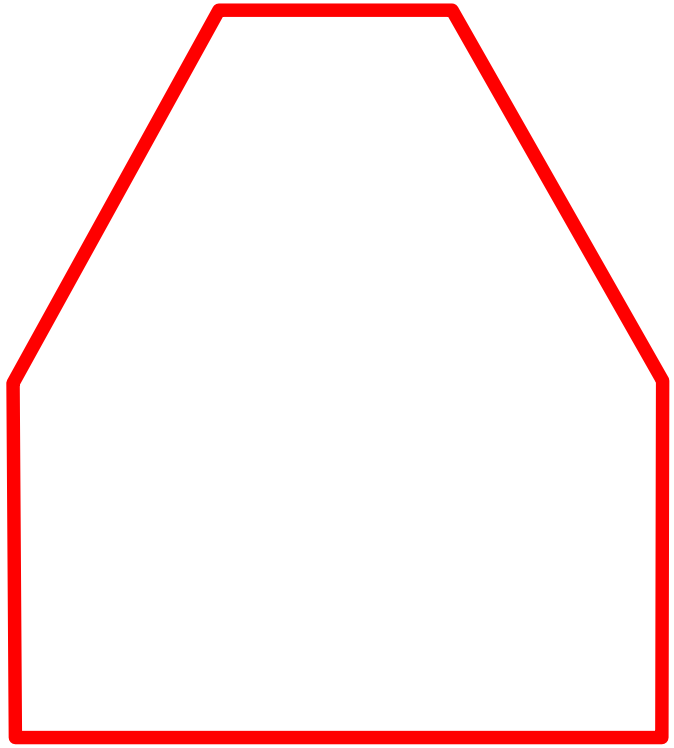
DNA as building block



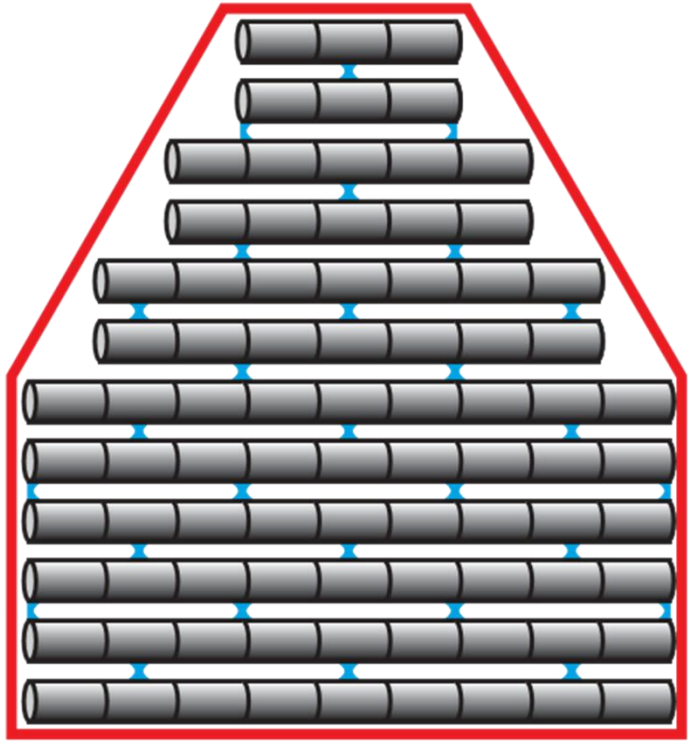
20 Å

36 Å

one turn, 10 ⅔ base pairs



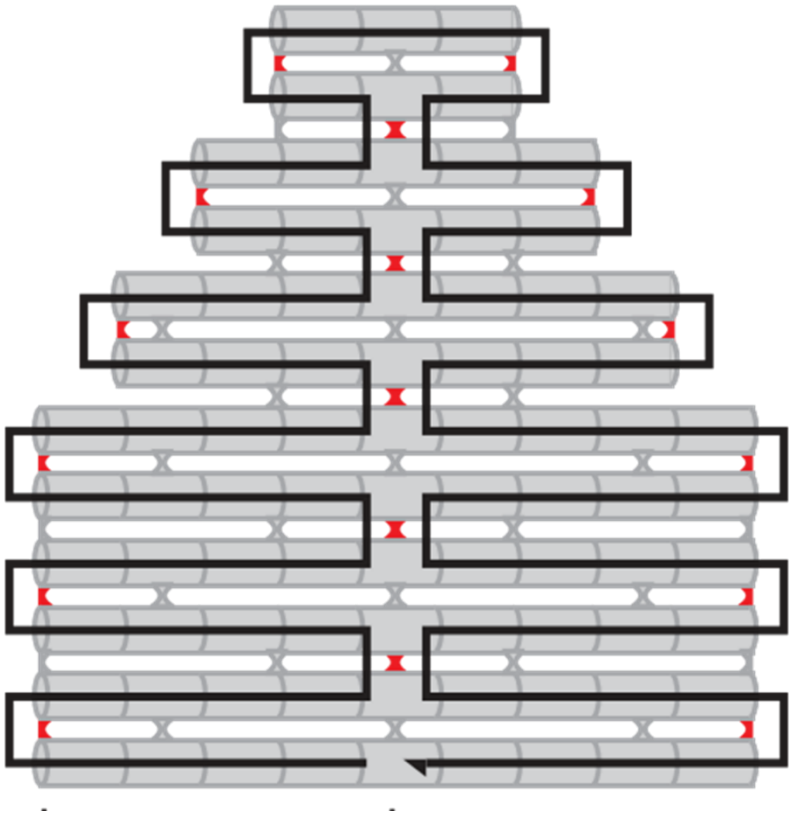
decide on shape

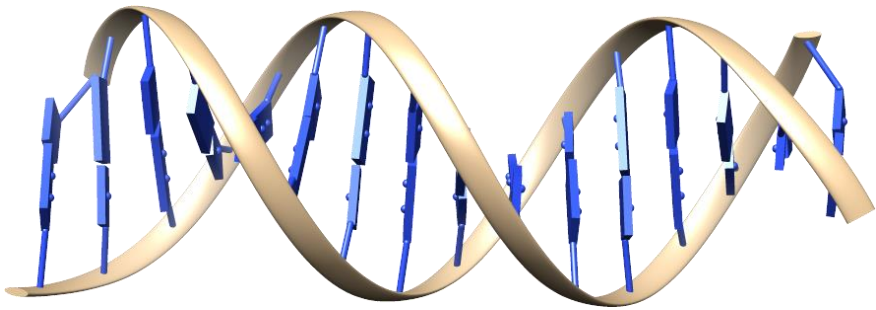


fill with cylinders
20 Å thick
length $\times \frac{10\ ⅔}{36}$ bases

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

One long strand runs along structure



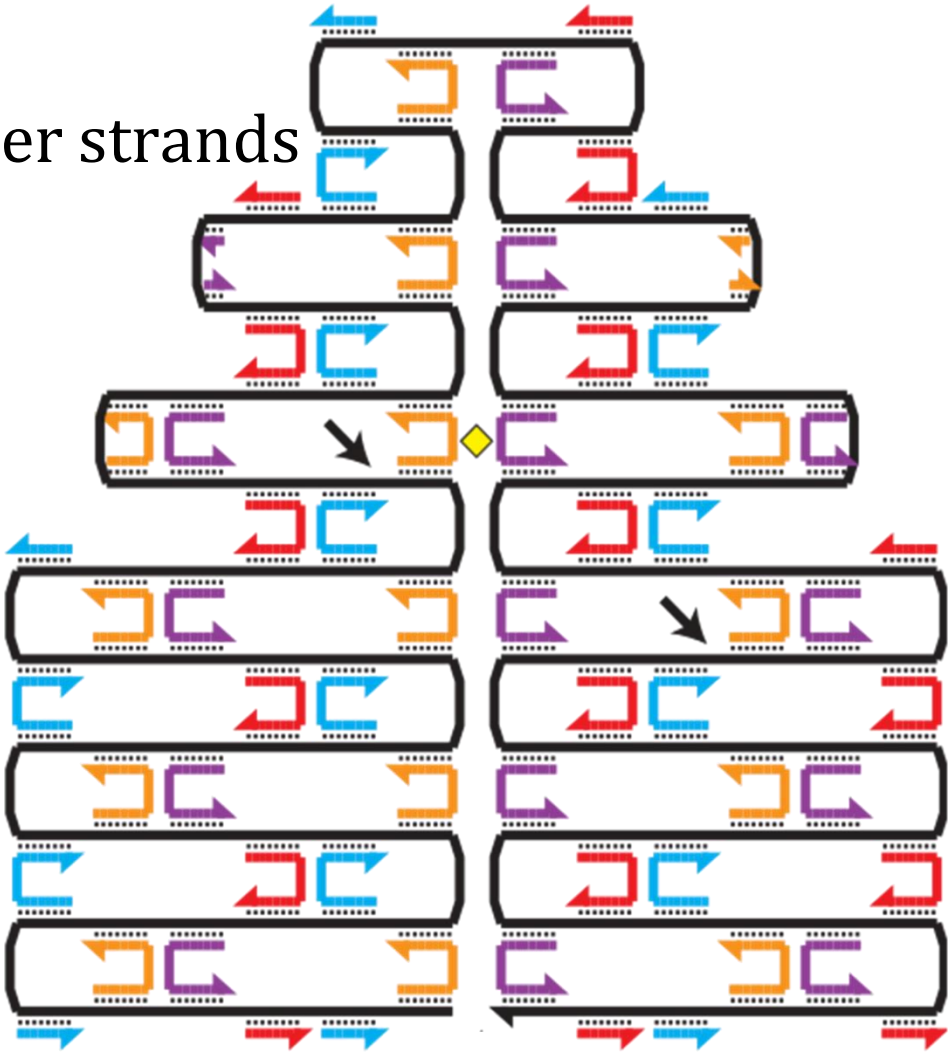


Every ½ turn brings other chain into position for crossing over…

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

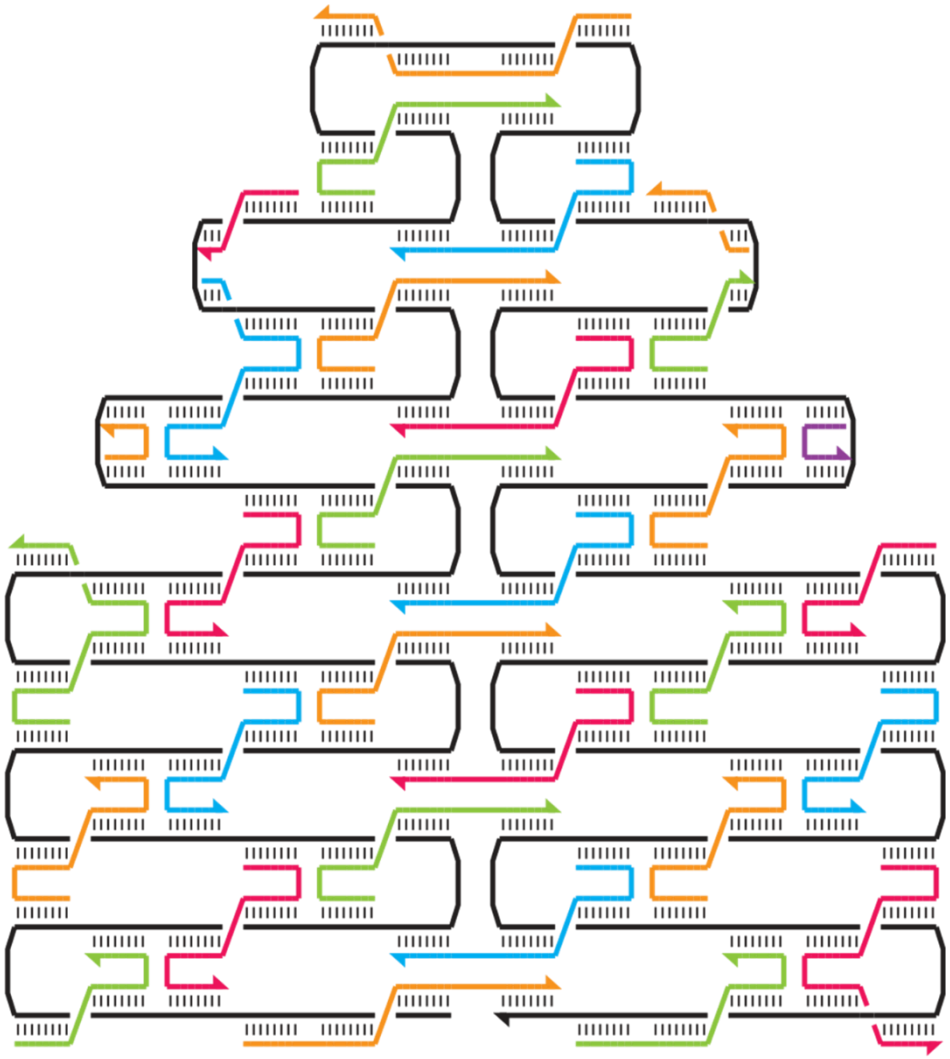
Decide where you would like “staples”

then join the staples into longer strands



result

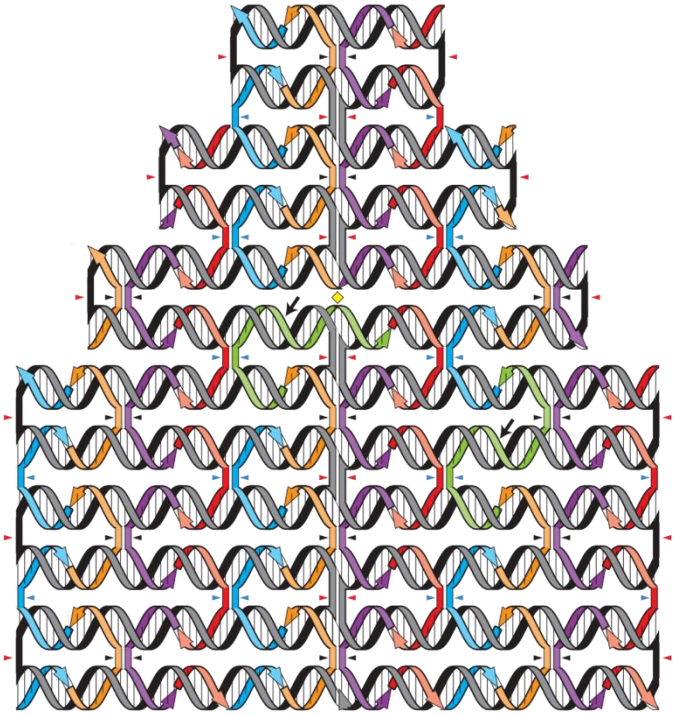
- every base is paired

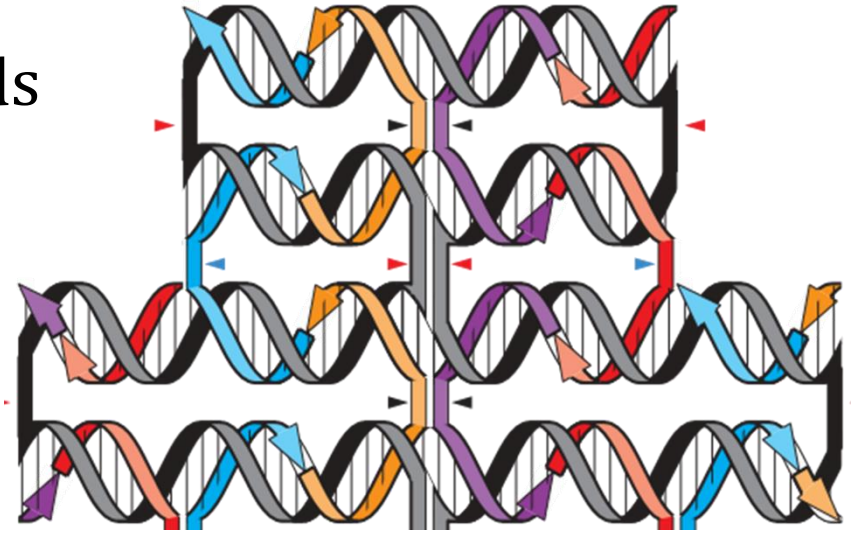


basically a long double helix

one long strand

lots of staple/joining strands





In this style of design

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

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

Where is the "negative design" ?

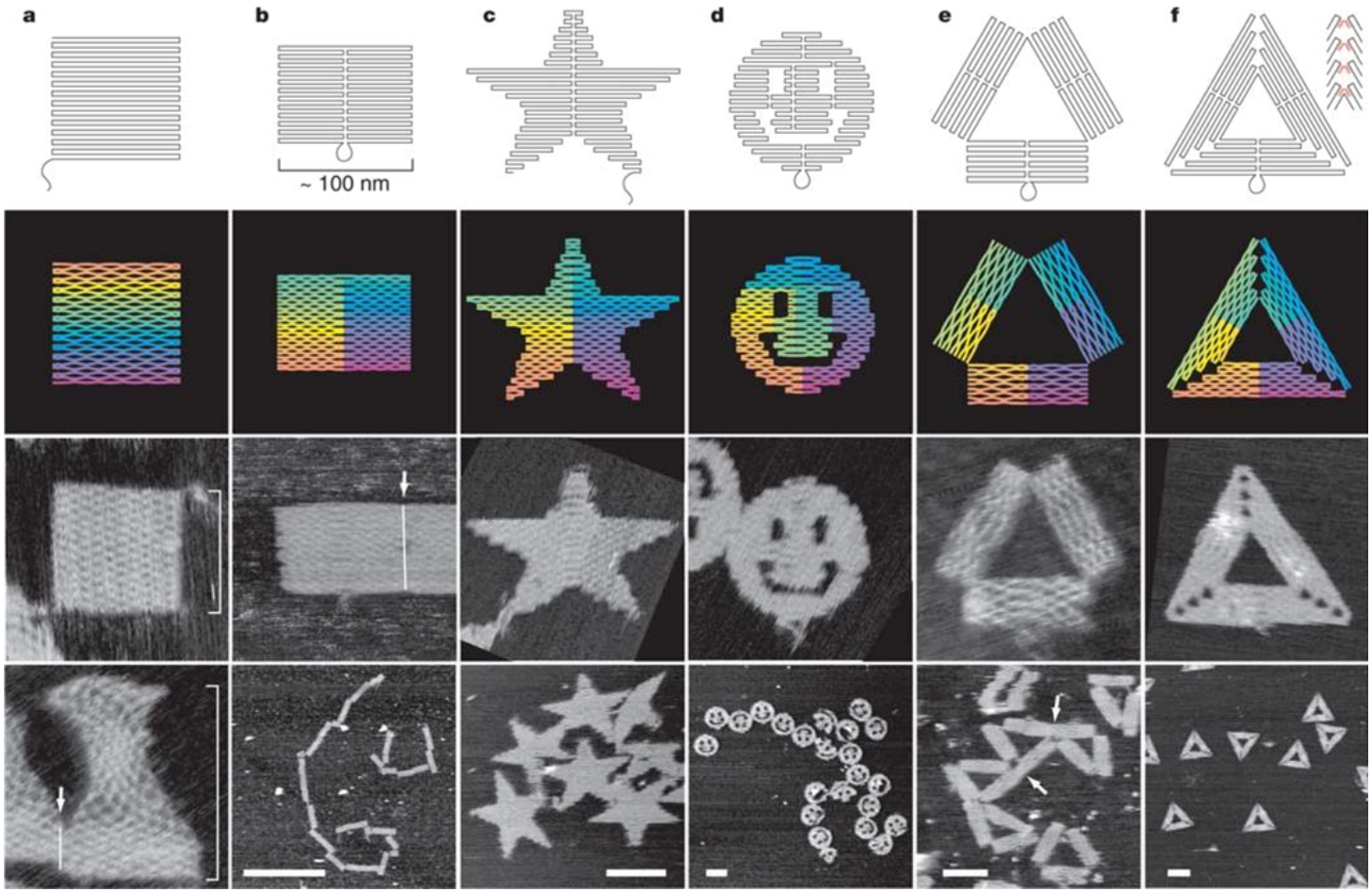
- you have a large natural piece of DNA – no repetitive elements
- staples fit to a specific part of long strand – not to other parts

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

spectacular success



designed shape

designed chain coloured

microscopy

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

compared to protein design

- much simpler energy approximations
  - DNA
    - just base-pairing
  - RNA
    - base pairs / stacking .. nearest neighbour energy model