Biopolymer design

Biopolymer design (inverse folding)

- Inverse folding of biopolymers: searching for sequences that fold into a given target structure.
- Inspired by the idea to design and exploit the molecules performing certain functions.
- Relevant for both RNA and proteins.

- Hard computational problem: it is not sufficient to find a sequence with low free energy when folded into the target shape. It is necessary to ensure the absence of other conformations with even lower free energies.
- Multiple different solutions are possible.

Inverse folding of RNA (secondary) structures

Usually is done by a stochastic procedure.

An algorithm starts from an initial seed sequence and improves it by an optimization algorithm.

For instance, the following main steps (Busch & Backofen, 2007):

Step 1: Finding a sequence that has the lowest free energy an RNA may have if folded into the target structure (dynamic programming). However, this initial sequence may have a minimum free energy structure that differs from the target.

Step 2: Iterative process of mutations that intend to reach a sequence with the free energy minimum in the target structure. For any intermediate solution, a “structural distance” between the lowest free energy conformation and the target structure is estimated.

Structural distance: e.g. the number of unique base pairs in each of the structures.

Inverse folding of RNA (secondary) structures

Usually is done by a stochastic procedure.

An algorithm starts from an initial seed sequence and improves it by an optimization algorithm.

For instance, the following main steps (Busch & Backofen, 2007):

Step 1: Finding a sequence that has the lowest free energy an RNA may have if folded into the target structure (dynamic programming). However, this initial sequence may have a minimum free energy structure that differs from the target.

Step 2: Iterative process of mutations that intend to reach a sequence with the free energy minimum in the target structure. For any intermediate solution, a “structural distance” between the lowest free energy conformation and the target structure is estimated.

Additional features in the RNA design programs:
- Certain sequence motif constraints.
- Optimization of the ensemble of structures (target folding frequency).

Structural distance: e.g. the number of unique base pairs in each of the structures.

Inverse folding of protein structures

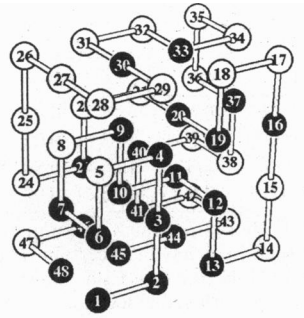
A stochastic procedure (e.g. Monte Carlo).
Difficult problem even in lattice-model simulations.

An interesting blind test of lattice-model-based strategies was published by the groups from Harvard University and University of California in San Francisco (UCSF) in 1995 (Yue et al.):

- The Harvard team used their inverse folding algorithm to design HP 48-mer sequences that should fold to 10 selected compact conformations.
- The sequences, but not the structures, were sent to the UCSF group.
- The UCSF group predicted the globally optimal conformations for the sequences.
- The structures were compared (lower energies of UCSF-predicted structures would mean a failure of the Harvard inverse folding).

HP cubic lattice model:

- Two monomer types: H, hydrophobic and P, polar.
- Energy: $E = -\varepsilon \times h$, where h is the number of H-H contacts between monomers that are not sequence neighbors, ε is a constant.



Inverse folding of protein structures

A stochastic procedure (e.g. Monte Carlo).
Difficult problem even in lattice-model simulations.

An interesting blind test of lattice-model-based strategies was published by the groups from Harvard University and University of California in San Francisco (UCSF) in 1995 (Yue et al.):

- The Harvard team used their inverse folding algorithm to design HP 48-mer sequences that should fold to 10 selected compact conformations.
- The sequences, but not the structures, were sent to the UCSF group.
- The UCSF group predicted the globally optimal conformations for the sequences.
- The structures were compared (lower energies of UCSF-predicted structures would mean a failure of the Harvard inverse folding).

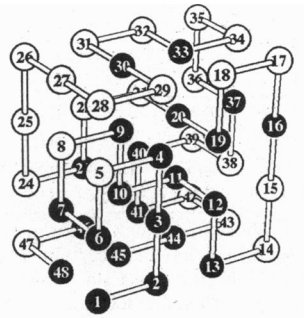
Yue et al.:

The Harvard group offered a six pack of beer if the UCSF group could successfully fold 1 or more out of 10 sequences of 48-mers. This is a report of the results.

Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 325–329, January 1995

HP cubic lattice model:

- Two monomer types: H, hydrophobic and P, polar.
- Energy: $E = -\varepsilon \times h$, where h is the number of H-H contacts between monomers that are not sequence neighbors, ε is a constant.

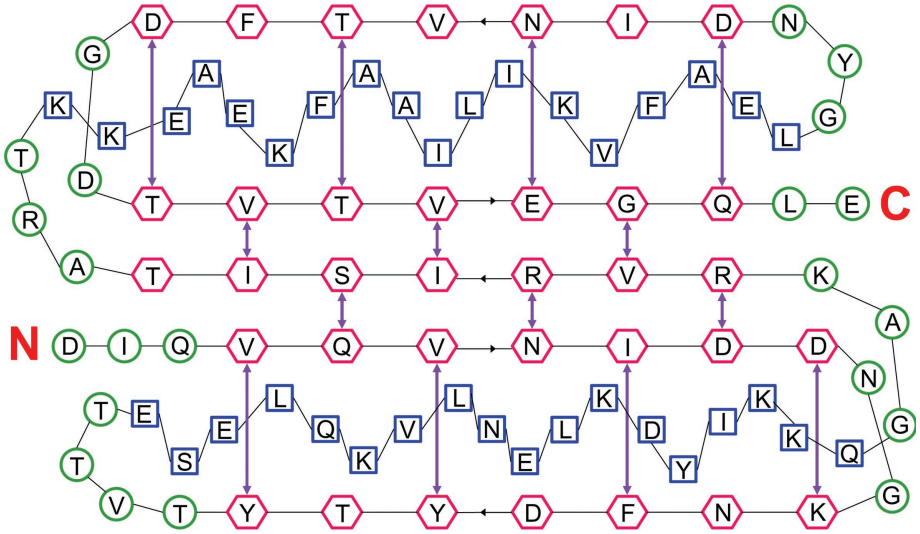


In 9 out of 10 sequences, the UCSF group found lower energy conformations than the 48-mers were designed to fold to.

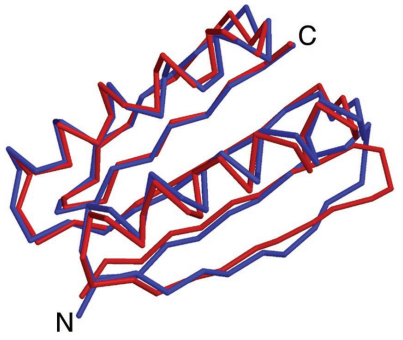
Inverse folding of protein structures

For a long time, the only successful examples were relatively simple coiled coils. However, it is possible to use e.g. an iterative procedure (structure prediction/ sequence optimization, back and forwards) to design a desired topology.

For instance, using Rosetta structure prediction protocols, a 93-residue protein with a desired topology was designed (Kuhlman et al., 2003).
Starting from initial “rough” structure, 15 cycles of sequence design and backbone optimization:



Desired “rough” topology

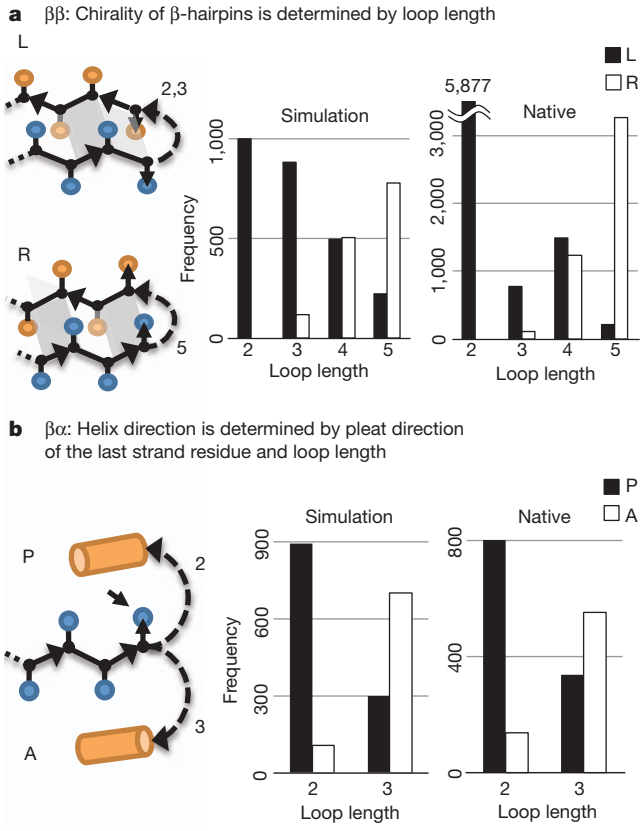


C-α overlay of the computed model (blue) and the solved X-ray structure (red)

Protein design

Some general principles (rules) can be used for the construction of local structures.

The conformations at the junctions of the adjacent secondary structure elements (e.g. ββ, βα, αβ) follow fundamental rules. For instance:



(Koga et al., 2012)

Protein design

It can be reasoned that evolution of natural proteins explored only a small region of the sequence space.

The number of amino acid sequences with typical protein length: 20^{200}.
The number of distinct proteins in extant organisms: ~10^{12}.

(Huang et al., 2016)

- Protein design can be based on an energy function that takes into account the interactions of the atoms in proteins with each other and with the solvent molecules. *The same function as the one used in protein structure prediction.*

- In the protein design, both the amino acid sequence and the conformational states of amino acid side chains are unknown.

- In the general *de novo* design, both the sequence and the conformations of the backbone and the side chains should be found.

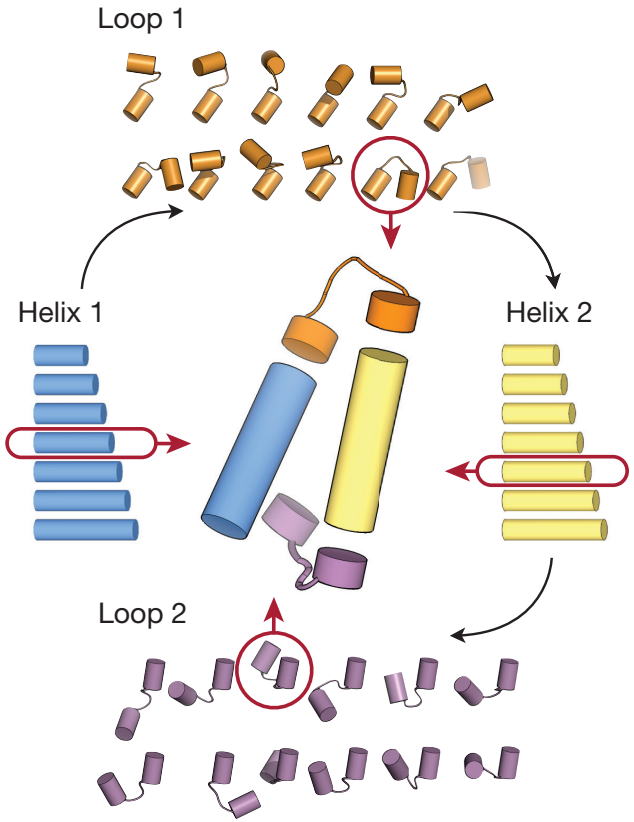
- The sequence optimization identifies the lowest-energy sequence for a given structure, and structure prediction checks whether the target structure is the lowest-energy conformation of the designed sequence.

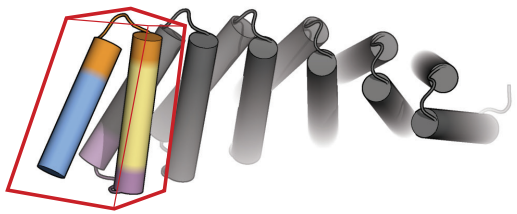
- Computational complexity of the combination of these two tasks can be reduced by general constraints in protein building blocks.

Protein design

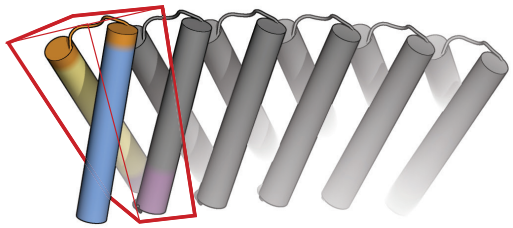
Repeat proteins

E.g. repeating structures with helix-loop-helix-loop combinations were constructed using Rosetta Monte-Carlo fragment assembly:





Right-handed repeat unit
16–3–17–3

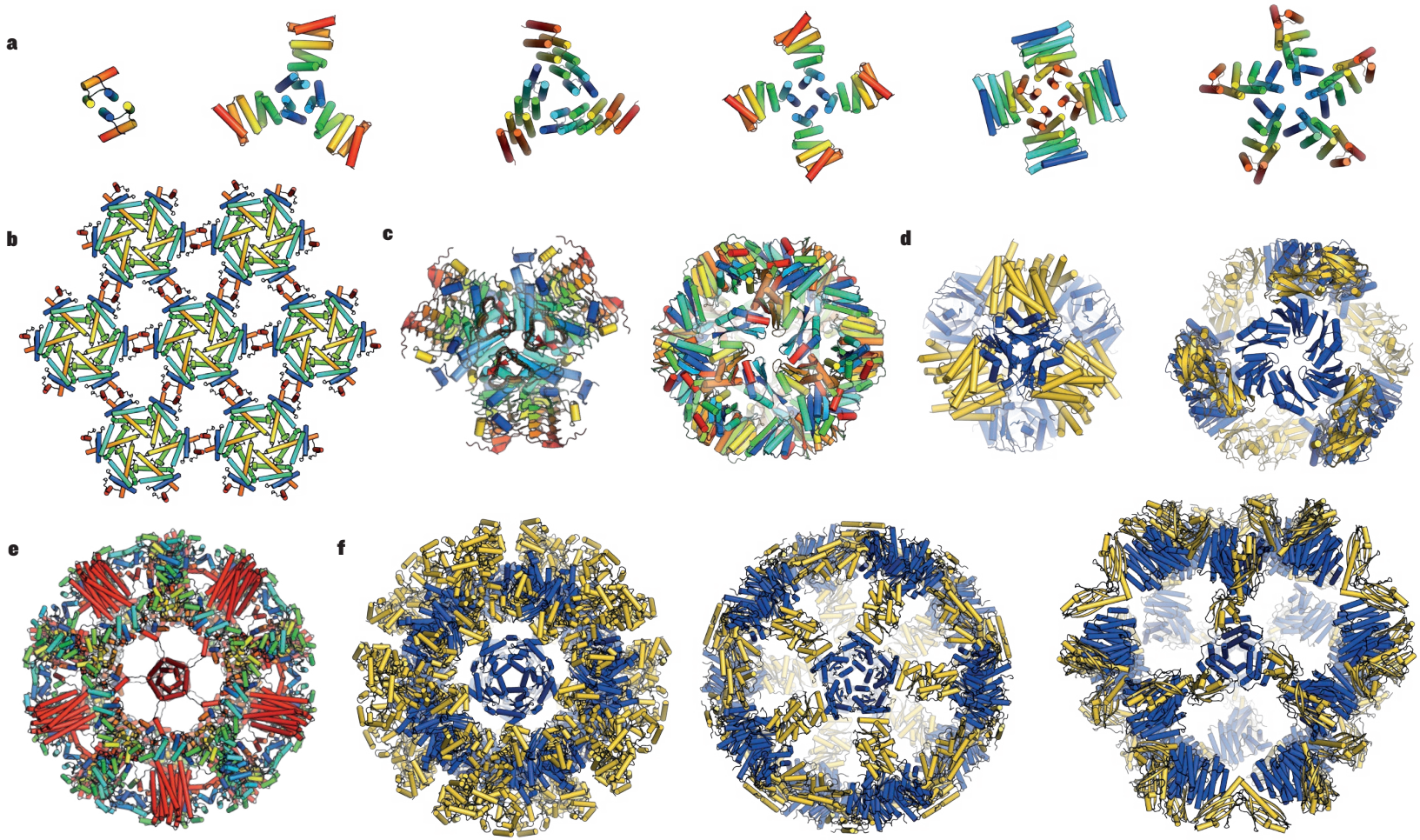


Left-handed repeat unit
22–4–20–4

(Brunette et al., 2015)

Protein design

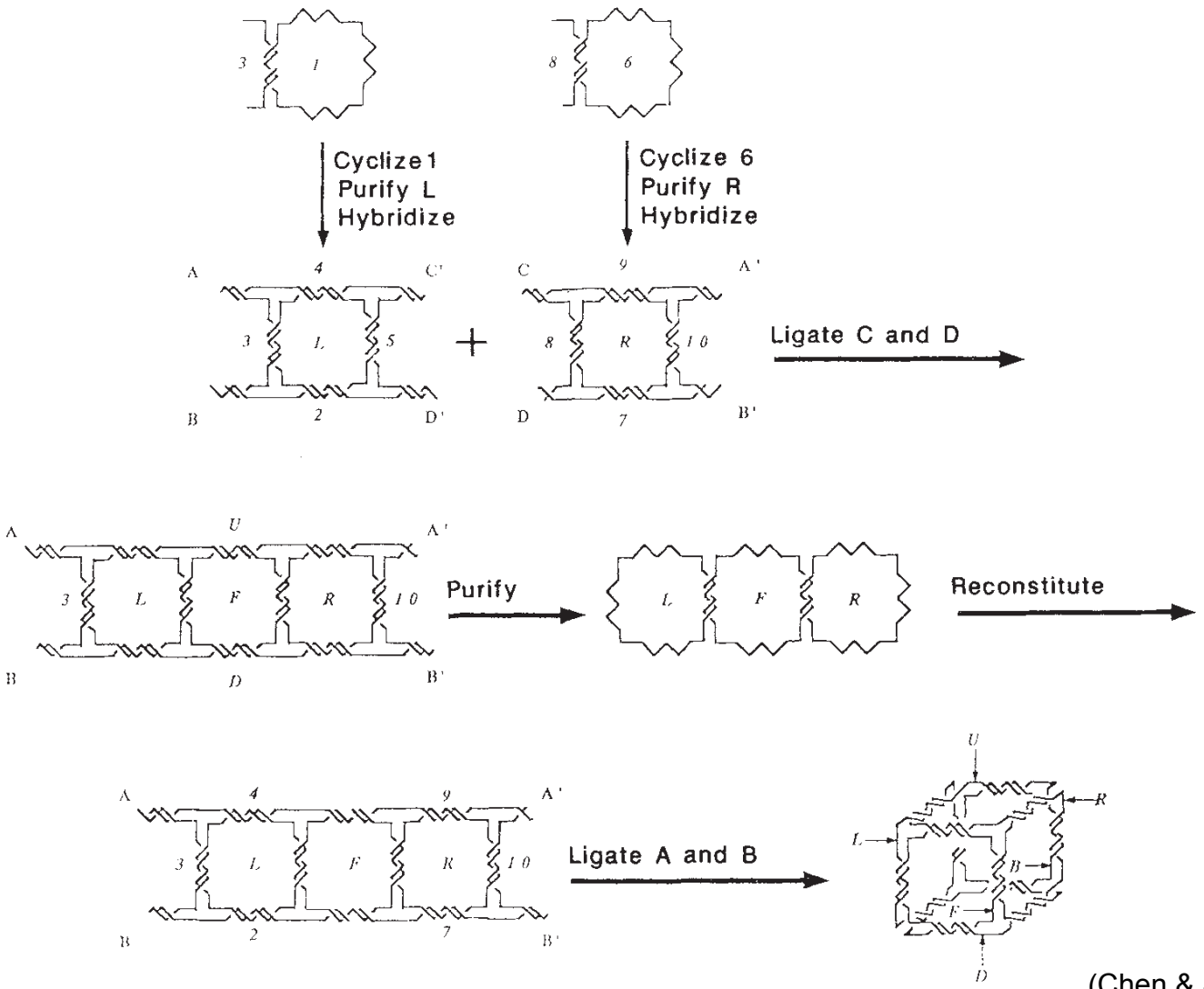
Design of new functions, e.g. self-assembling nanomaterials



from Huang et al. (2016)

DNA and RNA origami

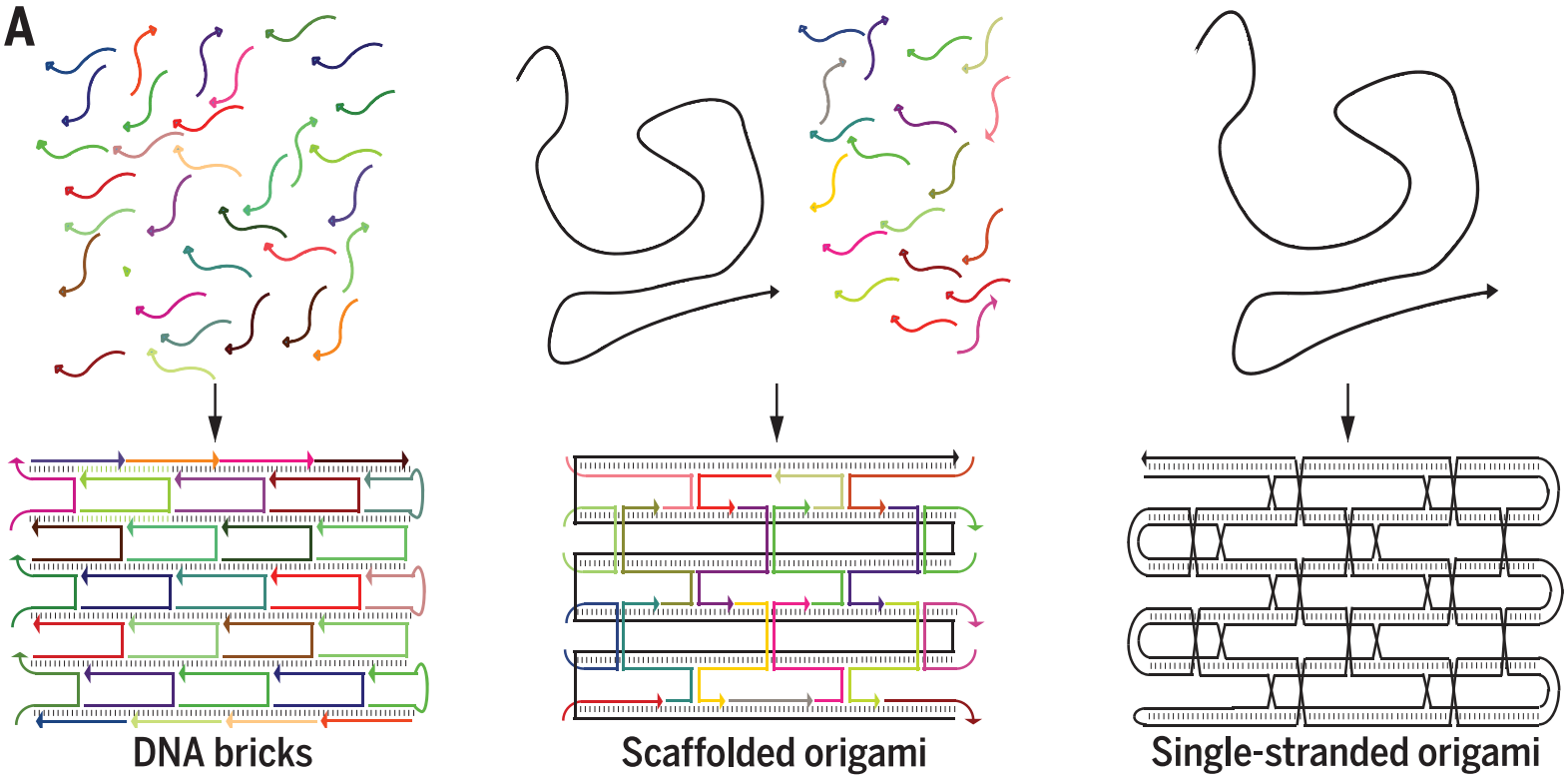
E.g. construction of a DNA cube:



(Chen & Seeman, 1991)

DNA and RNA origami

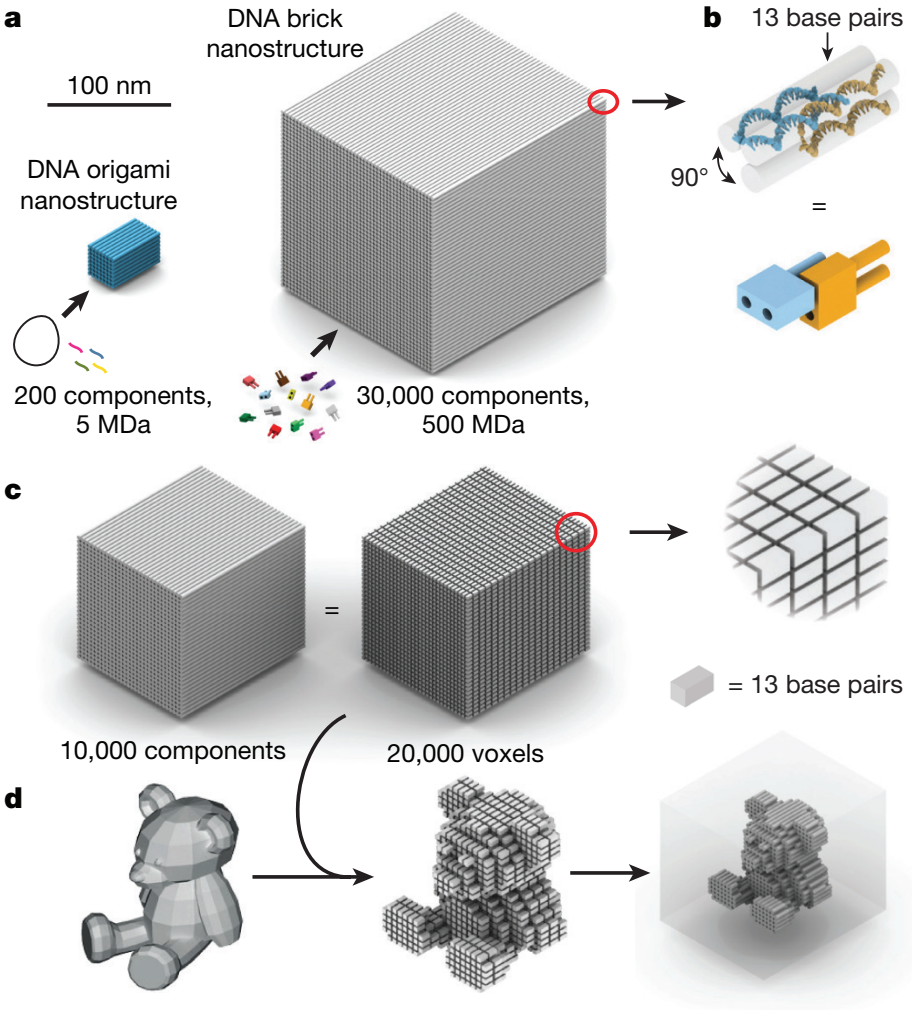
Various nanostructures can be constructed in a programmable way from multiple blocks



(Han et al., 2017)

DNA and RNA origami

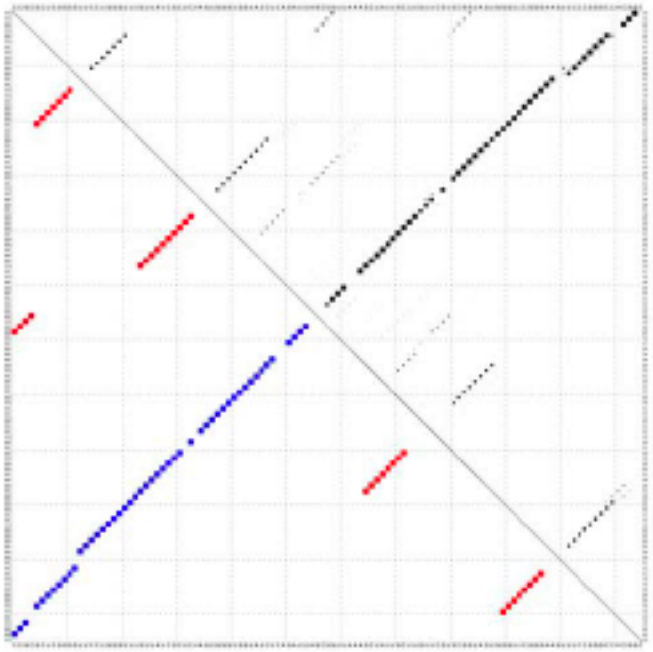
Various nanostructures can be constructed in a programmable way from multiple blocks

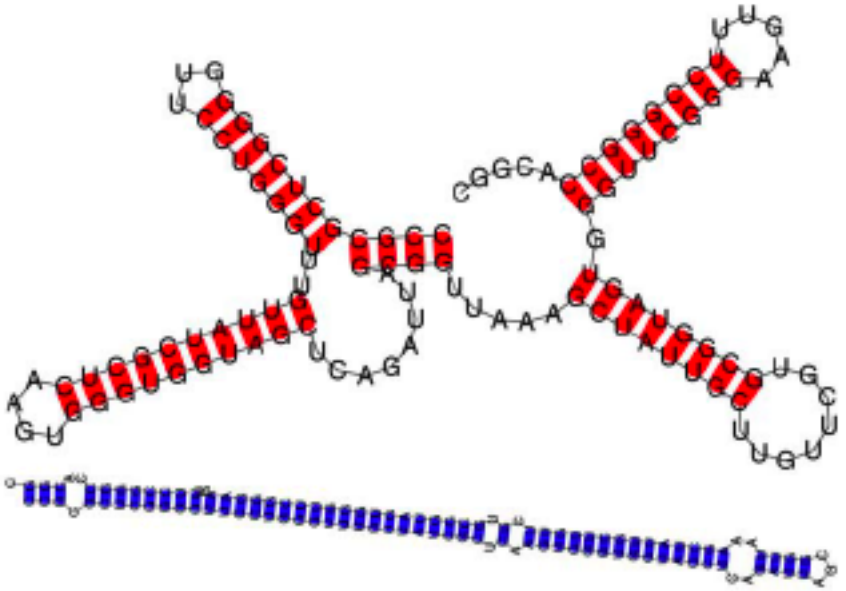


(Ong et al., 2017)

Design of bistable RNA structures

- Design of RNA molecules with two different alternatives: a conformational switch.
- Much more complex optimization problem.
- Some algorithms were suggested.



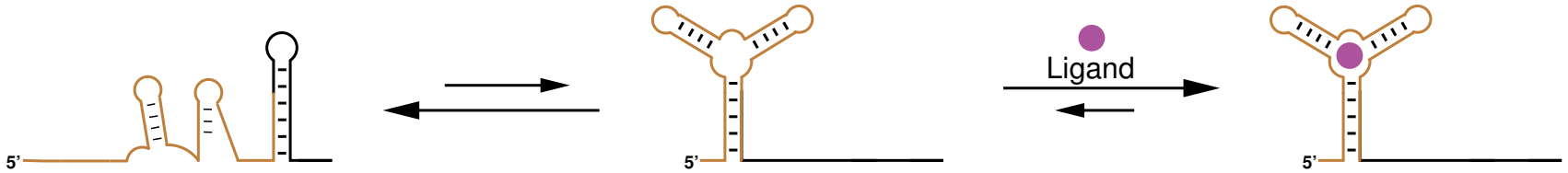


Höner zu Siederdissen et al. (2013)

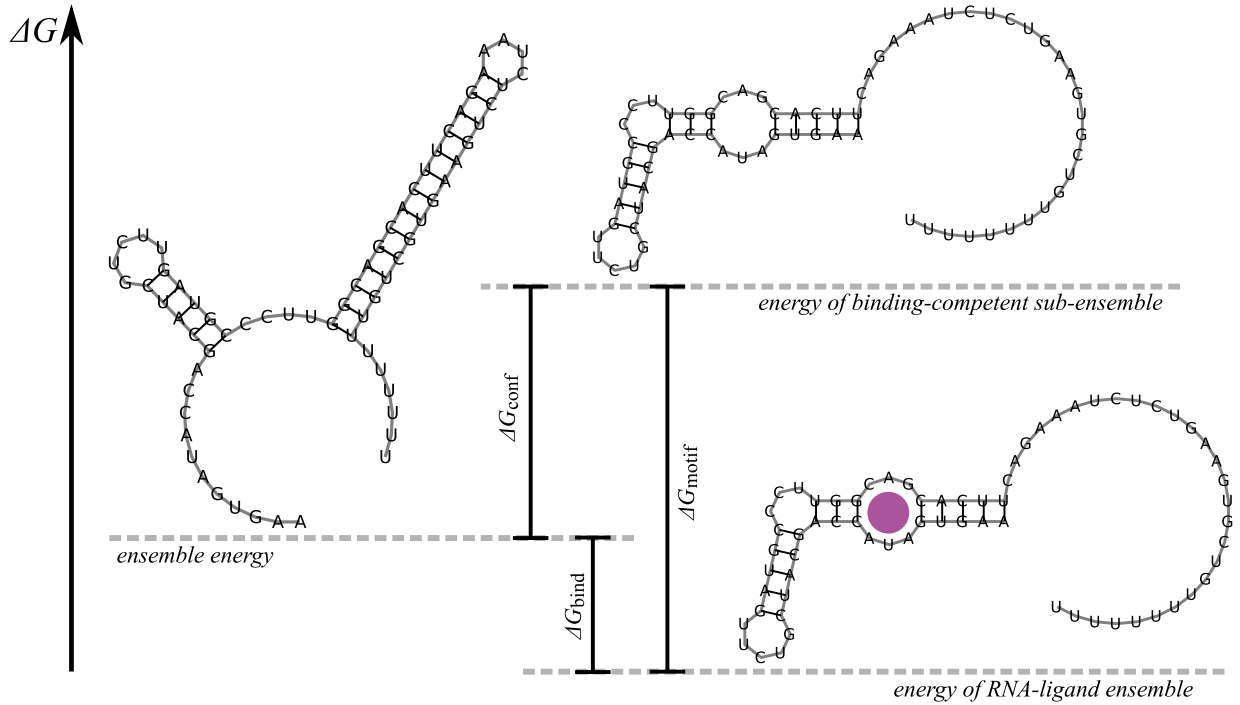
Design of riboswitches / RNA sensors

RNA elements that adopt alternative conformations can sense the presence of ligands.

Illustration of the general mechanism (Findeiß et al., 2017):



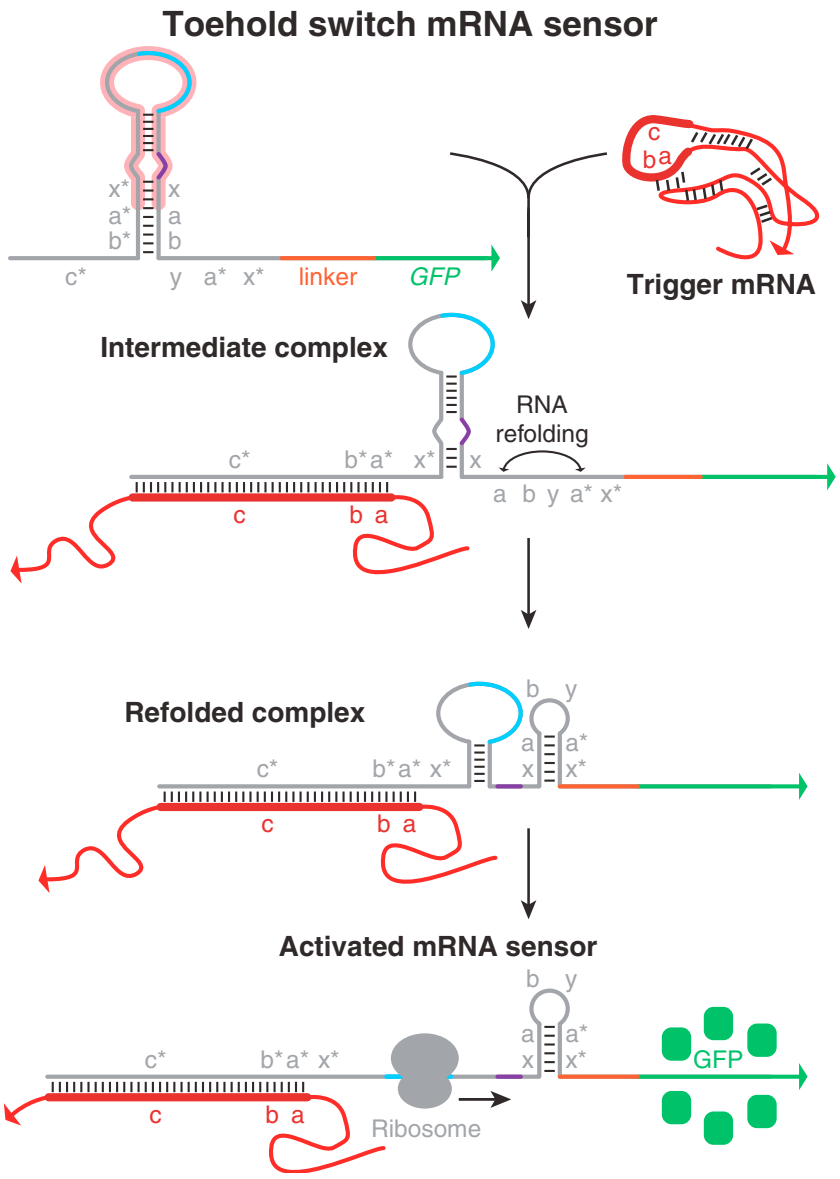
Thermodynamics and kinetics of alternative structures should be calculated with constraints imposed by ligand binding:



(From Findeiß et al., 2017)

Design of riboswitches / RNA sensors

- Toehold switches are designed to sense the presence of specific RNA, e.g. virus mRNA.
- The sensor contains a reporter gene, e.g. the green fluorescent protein (GFP).
- The GFP expression is suppressed by structures prohibiting ribosome binding.
- The conformational switch triggered by the sensed RNA binding releases the GFP translation.



(Green et al., 2014)