STOCHASTIC LOCAL SEARCH
FOUNDATIONS AND APPLICATIONS

DNA Code Design

Holger H. Hoos & Thomas Stützle

Introduction

DNAs, RNAs, Proteins

- crucial components of biochemistry of all organisms
- chemical structure: chain of simple building blocks
- sequence representation: strings over (small) alphabet

The Central Dogma of Molecular Biology

DNA ⟶ mRNA ⟶ protein

- genetic information is stored as double stranded DNA
- DNA is transcribed into messenger RNA (mRNA)
- mRNA is translated into proteins
  by ribosomes, using tRNAs for decoding
- proteins catalyse biochemical reactions,
  serve as structural building blocks

DNA Code Design

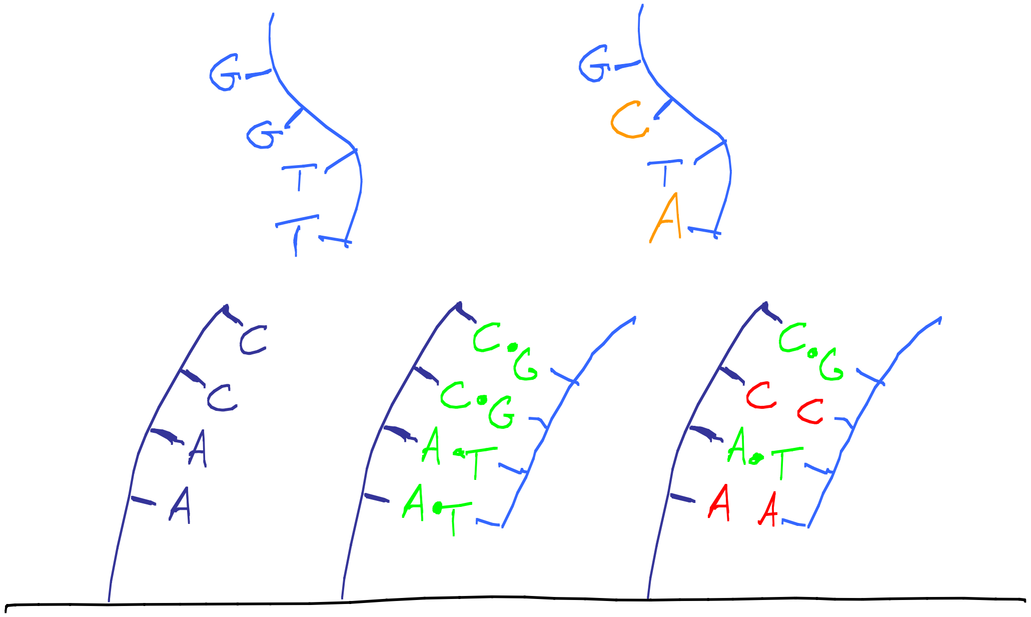
[joint work with Dan Tulpan and Anne Condon; in collaboration
with Seo Chang Bong, Michael Shortread, and Lloyd Smith]

Goal: Design sets of DNA strands satisfying given constraints

Applications:

- DNA computing
- DNA nanostructure design
- biomolecular tagging
- DNA microarray design

Example: Surfaced-Based DNA Computing



Constraints:

- Two strings in set S must be either perfect complements or sufficiently different

$\leadsto$ *Hamming distance constraints:*

HD(d): Any two different strings in S must differ in at least d positions.

CHD(d): Any string s in S must from the complement of any other string s' differ in at least d positions.

Constraints:

- Two strings in set S must be either perfect complements or sufficiently different

$\rightsquigarrow$ *Hamming distance constraints:*

HD(d): Any two different strings in S must differ in at least d positions.

CHD(d): Any string s in S must from the complement of any other string s' differ in at least d positions.

- Uniform melting temperature of correct double-stranded hybrids

$\rightsquigarrow$ *GC content constraint:*

GCC(p): For any string in S, a fraction p of its positions are Gs and Cs.

Nomenclature:

- DNA strands = code words
- set of strands = code
- number of strands in a set = size of code

DNA Code Design Problem:

Given one or more design constraints, a word length n, and target number of code words m, find a code of size m that satisfies the constraints.

DNA code design problems are ...

- related to problems in coding theory (robust codes),
- (believed to be) computationally hard,
- not efficiently soluble in practice.

Solution approach:

Use Stochastic Local Search Method for constructing word sets satisfying given constraints:

initialisation: create set of words by random construction

search steps: pick conflict, mutate one of the words involved such that conflict is reduced (greedy randomised mechanism)

termination: end search when satisfying word set is found, or maximal number of iterations have been performed without finding a solution

Sample DNA Code: HD(4),CHD(4),GCC(0.5), word length 8, size 112

```
AACCACCA
TAGCAGCT
TAGCTCGA
ATCGAGCT
ATCGTCGA
AAGGTGCA
AAGGACGT
TTCCTGCA
TTCCACGT
TACGTCCT
TACGAGGA
ATGCTCCT
ATGCAGGA
TTGGACCA
AACCTGGT
TTGGTGGT
ACCAAACC
TCGAATCG
TCGATAGC
AGCTATCG
AGCTTAGC
ACGTTTCC
ACGTAAGG
TGCATTCC
TGCAAAGG
TCCTTACG
TCCTATGC
AGGATACG
AGGAATGC
TGGTAACC
ACCATTGG
TGGTTTGG
CCAACAAC
GCTACTAG
GCTAGATC
CGATCTAG
CGATGATC
CCTTGTAC
CCTTCATG
GGAAGTAC
GGAACATG
GCATGAAG
GCATCTTC
CGTAGAAG
CGTACTTC
GGTTCAAC
CCAAGTTG
GGTTGTTG
CAACCCAA
GATCCGAT
GATCGCTA
CTAGCGAT
CTAGGCTA
CATGGGAA
CATGCCTT
GTACGGAA
GTACCCTT
GAAGGCAT
GAAGCGTA
CTTCGCAT
CTTCCGTA
GTTGCCAA
CAACGGTT
GTTGGGTT
ACACACAC
TCTCAGAG
TCTCTCTC
AGAGAGAG
AGAGTCTC
ACTGTGAC
ACTGACTG
TGACTGAC
TGACACTG
TCAGTCAG
TCAGAGTC
AGTCTCAG
AGTCAGTC
TGTGACAC
ACACTGTG
TGTGTGTG
AAAACCCC
TATACGCG
TATAGCGC
ATATCGCG
ATATGCGC
AATTGGCC
AATTCCGG
TTAAGGCC
TTAACCGG
TAATGCCG
TAATCGGC
ATTAGCCG
ATTACGGC
TTTTCCCC
AAAAGGGG
TTTTGGGG
CACAAGAG
GAGTAGTG
GTGTTCTC
GTCTACAG
CAGATGTC
CAGTTCAG
ACTGGTGT
AGACGTGA
AGTCCTCT
TCTGCACT
TGTCGAGT
TCAGGAGA
GTCAAGTC
GTGATGAG
TGTGGTCA
ACAGCTCA
```

Sample DNA Code: HD(4),CHD(4),GCC(0.5), word length 8, size 112

```
AACCACCA
TAGCAGCT
TAGCTCGA
ATCGAGCT
ATCGTCGA
AAGGTGCA
AAGGACGT
TTCCTGCA
TTCCACGT
TACGTCCT
TACGAGGA
ATGCTCCT
ATGCAGGA
TTGGACCA
AACCTGGT
TTGGTGGT
ACCAAACC
TCGAATCG
TCGATAGC
AGCTATCG
AGCTTAGC
ACGTTTCC
ACGTAAGG
TGCATTCC
TGCAAAGG
TCCTTACG
TCCTATGC
AGGATACG
AGGAATGC
TGGTAACC
ACCATTGG
TGGTTTGG
CCAACAAC
GCTACTAG
GCTAGATC
CGATCTAG
CGATGATC
CCTTGTAC
CCTTCATG
GGAAGTAC
GGAACATG
GCATGAAG
GCATCTTC
CGTAGAAG
CGTACTTC
GGTTCAAC
CCAAGTTG
GGTTGTTG
CAACCCAA
GATCCGAT
GATCGCTA
CTAGCGAT
CTAGGCTA
CATGGGAA
CATGCCTT
GTACGGAA
GTACCCTT
GAAGGCAT
GAAGCGTA
CTTCGCAT
CTTCCGTA
GTTGCCAA
CAACGGTT
GTTGGGTT
ACACACAC
TCTCAGAG
TCTCTCTC
AGAGAGAG
AGAGTCTC
ACTGTGAC
ACTGACTG
TGACTGAC
TGACACTG
TCAGTCAG
TCAGAGTC
AGTCTCAG
AGTCAGTC
TGTGACAC
ACACTGTG
TGTGTGTG
AAAACCCC
TATACGCG
TATAGCGC
ATATCGCG
ATATGCGC
AATTGGCC
AATTCCGG
TTAAGGCC
TTAACCGG
TAATGCCG
TAATCGGC
ATTAGCCG
ATTACGGC
TTTTCCCC
AAAAGGGG
TTTTGGGG
CACAAGAG
GAGTAGTG
GTGTTCTC
GTCTACAG
CAGATGTC
CAGTTCAG
ACTGGTGT
AGACGTGA
AGTCCTCT
TCTGCACT
TGTCGAGT
TCAGGAGA
GTCAAGTC
GTGATGAG
TGTGGTCA
ACAGCTCA
```

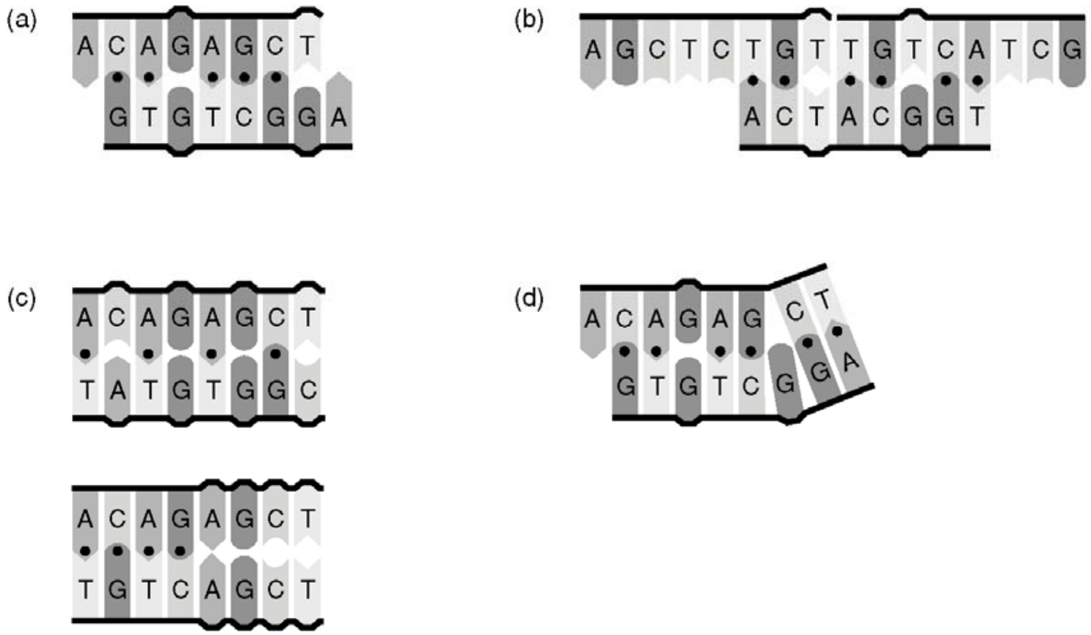
New Empirical Bounds on DNA Word Sets for HD and CHD Constraints

$n \setminus d$	4	5	6	7	8	9	10
4	**2** $[.02k]$	-	-	-	-	-	-
5	**4** $[.02k]$	**2** $[.005k]$	-	-	-	-	-
6	**28** $[28k]$	**4** $[.1k]$	**2** $[.02k]$	-	-	-	-
7	**40** $[254k]$	**11** $[3k]$	**2** $[.005k]$	**2** $[.05k]$	-	-	-
8	**112** $[850k]$	**27** $[7k]$	**10** $[31k]$	**2** $[.01k]$	**2** $[.03k]$	-	-
9	**314** $[677k]$	**72** $[142k]$	**20** $[37k]$	**8** $[16k]$	**2** $[.02k]$	**2** $[.04k]$	-
10	**938** $[702k]$	**180** $[386k]$	**49** $[287k]$	**16** $[495k]$	**8** $[11.01k]$	**2** $[.01k]$	**2** $[.02k]$
11	**2750** $[117k]$	**488** $[257k]$	**114** $[145k]$	**35** $[6k]$	**12** $[1k]$	**5** $[2k]$	**2** $[.05k]$
12	$>$**8000** $[72k]$	**1340** $[400k]$	**290** $[327k]$	**79** $[236k]$	**27** $[712.5k]$	**11** $[828k]$	**4** $[.2k]$

Results:

- large number of word sets exceeding best known theoretical constructions for various combination of constraints
- promising results for related problems from classical coding theory (binary / quaternary Hamming codes)
- promising results for DNA word design with more realistic, thermodynamic constraints

Various Forms of DNA-DNA Hybridisation



Combinatorial constraints (*e.g.* HD, CHD, GCC) provide only rough approximation to interactions between DNA strands

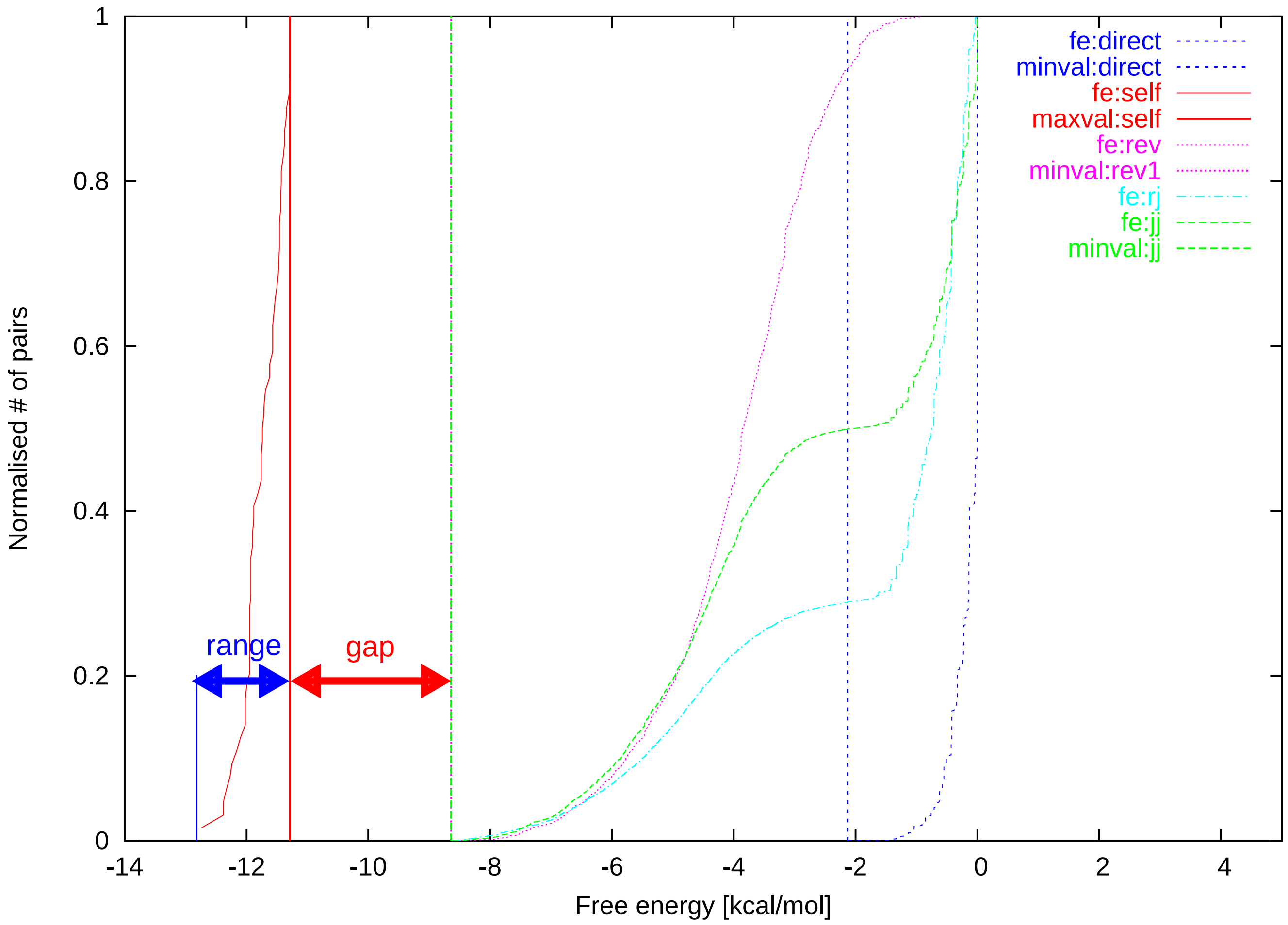
$\leadsto$ model interactions more precisely based on thermodynamics of DNA hybridisation

(Closely related to RNA structure prediction!)

DNA code design with thermodynamic constraints:

- model duplex stability using free energy calculation (or melting temperature prediction)
- choose thermodynamic constraints, *e.g.*, bound on energy gap between correct and incorrect hybridisations
- use extended version of SLS algorithm for combinatorial constraints to construct codes satisfying such constraints
  $\leadsto$ excellent empirical results

Example of a Thermodynamically Well-Behaved Code



Computational Results from Thermodynamic Code Design

	Original Codes from Literature			New Codes		
Strand	**Number of**	**melting temp**	**mfe gap**	**Number of**	**melting temp**	**mfe gap**
Length	**Strands**	**[°C]**	**[kcal/mol]**	**Strands**	**[°C]**	**[kcal/mol]**
8	108	12.41 . . . 22.47	-2.22	**200**	12.41 . . . 22.43	0.51
				300	12.48 . . . 22.43	0.42
15	40	43.79 . . . 52.74	4.28	**80**	46.66 . . . 52.67	5.80
				100	43.95 . . . 52.24	4.60
15	20	38.18 . . . 55.40	3.65	**40**	42.45 . . . 54.64	5.25
				60	43.01 . . . 46.87	5.28
				60	39.12 . . . 53.97	6.05
				60	47.59 . . . 52.67	6.63
16	24	52.37 . . . 55.75	8.12	**30**	53.06 . . . 53.97	8.75

Computational Results from Thermodynamic Code Design

	Original Codes from Literature			New Codes		
Strand Length	**Number of Strands**	**melting temp [°C]**	**mfe gap [kcal/mol]**	**Number of Strands**	**melting temp [°C]**	**mfe gap [kcal/mol]**
8	108	12.41 . . . 22.47	-2.22	**200**	12.41 . . . 22.43	0.51
				300	12.48 . . . 22.43	0.42
15	40	43.79 . . . 52.74	4.28	**80**	46.66 . . . 52.67	5.80
				100	43.95 . . . 52.24	4.60
15	20	38.18 . . . 55.40	3.65	**40**	42.45 . . . 54.64	5.25
				60	43.01 . . . 46.87	5.28
				60	39.12 . . . 53.97	6.05
				60	47.59 . . . 52.67	6.63
16	24	52.37 . . . 55.75	8.12	**30**	53.06 . . . 53.97	8.75