Universal DNA TAT Systems: Combinatorial Designs

EE 150, CALTECH, Fall 2004

Presentation by:

Mostafa El-Khamy

Main Reference

[0] Universal DNA Tag Systems: A Combinatorial Design Scheme, A. Ben-Dor, R. Karp, B. Schwikowski and Z. Yakhini

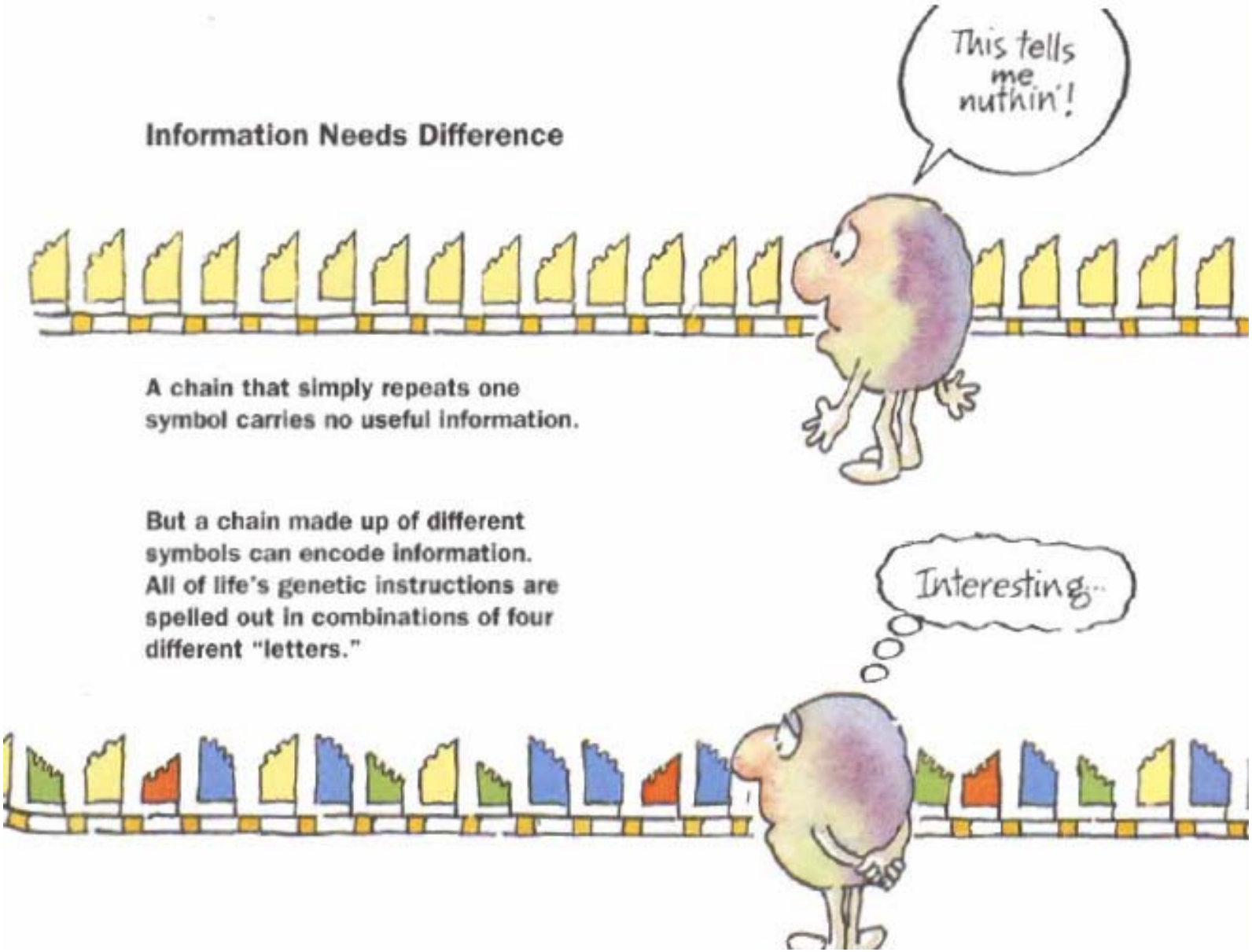
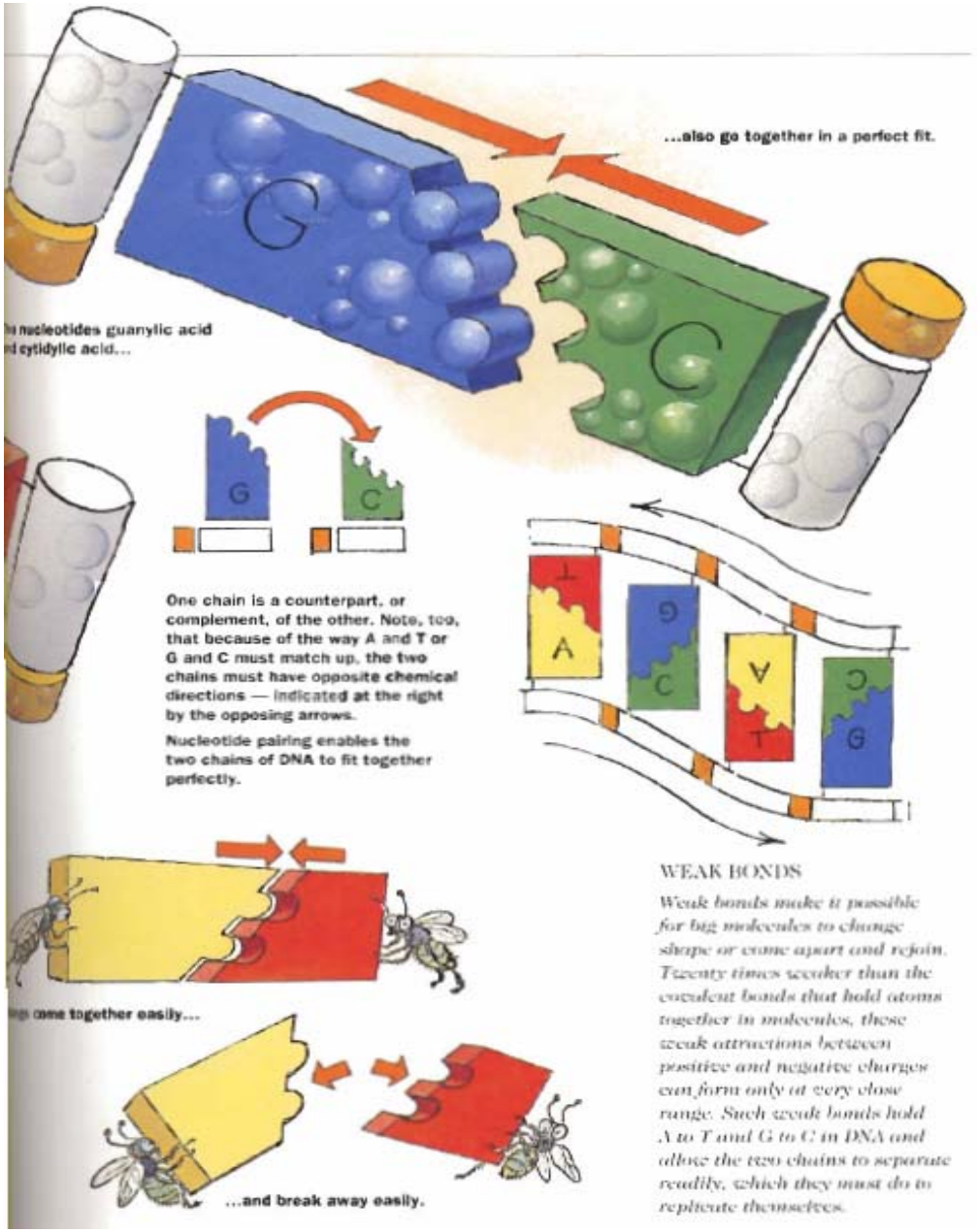


Figure from `The Way Life Works,'Mahlon Hoagland and Bert Dodson.

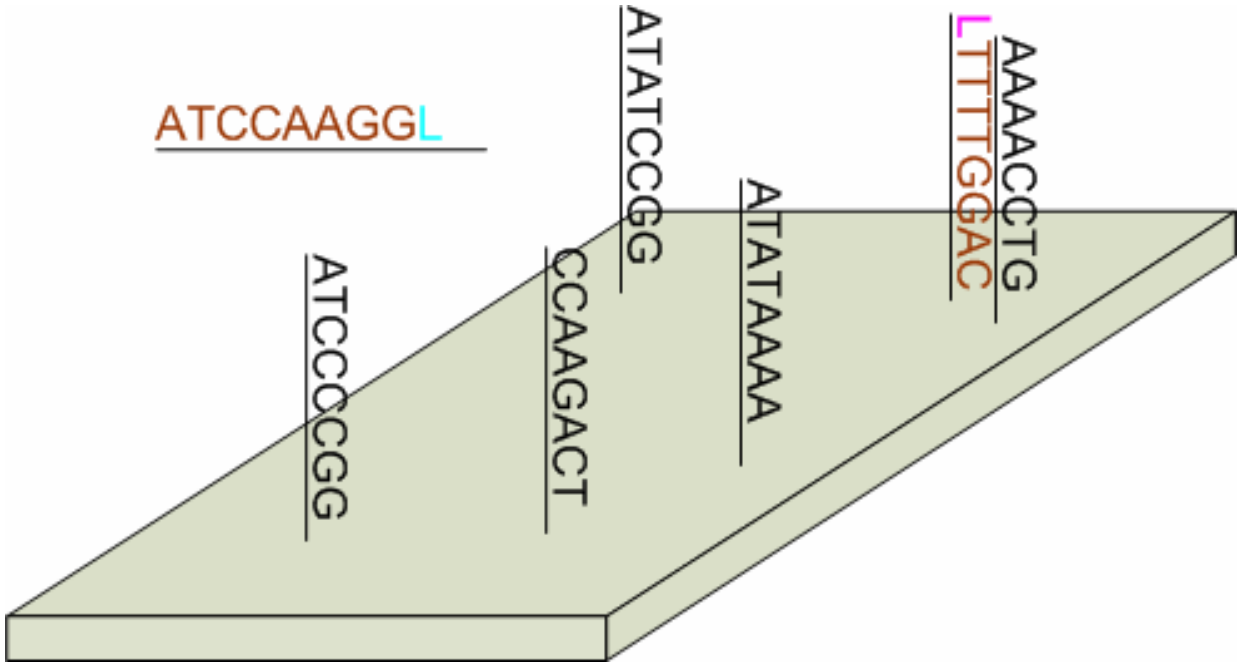


- A: Adenylic Acid
- T: Thymidylic Acid
- G: Guanylic Acid
- C: Cytidylic Acid
  - $A=T^{c}$
  - $C=G^{c}$

Figure from `The Way Life Works,'Mahlon Hoagland and Bert Dodson.

DNA —transcription→ mRNA —translation→ Protien

- Array Based Hybridization assays [6,13,15,..]
- Probe (Antitag): A target specific set of Oligonucleotides is immobilized to a substrate. (Microarray)
- Target sample of RNA or DNA is flourescently labeled (Tags)
- Brought in Contact with the Microarray and allowed to hybridize
- Information about the sample content is obtained by scanning the fluorescent labels



- "Yeast genome microarray. The actual size of the microarray is 18 mm by 18 mm.
- A fluorescently labeled cDNA probe was prepared from mRNA isolated from cells harvested shortly after

  inoculation
- Similarly, a second probe was prepared from mRNA isolated from cells taken from the same culture 9.5 hours later"

- **`Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale,' Joseph L. DeRisi, Vishwanath R. Iyer, Patrick O. Brown**

AFFYMETRIX GENECHIP(R) BRAND HUMAN GENOME U133 PLUS 2.0 ARRAY

Affymetrix News Release - 10/2/03

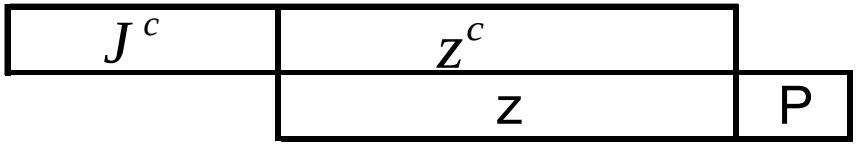
- "10,000 new probe sets representing about 6,500 new genes, for a total of nearly 50,000 RNA transcripts and variants. This new information, verified against the latest version of the publicly available genome map, provides researchers the most comprehensive and up-to-date genome-wide gene expression analysis."

Universal Microarrays

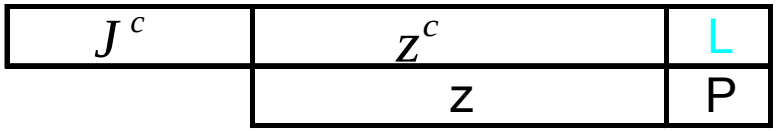
- Developed by Brenner [1], Morris et. al [2],…
- Example: SNP (Single Nucleotide Polymorphisms Detection)
  - AACCTT-> AACCTC

- Microarrays are efficient in detecting genetic variations in multiple samples simultaneously.
- A microarray is designed with a set of *universal* antitags, $\Omega=\{J_1, J_2,\ldots, J_n\}$.
- A unique Tag in Ω^c is ligated to a target specific complementary sequence to form a reporter molecule.

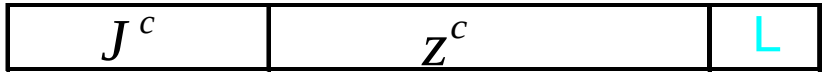
- Solution Phase: Hybridization between target DNA and reporter molecules.



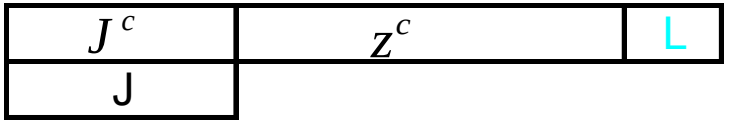
- In Polymerase driven reaction each hybridized reporter molecule is extended by a fluorescently labeled dideoxynucleotide. (ddA,ddT,ddC,ddG).



- Separate Extended Reporters from the Sample fragments.



- Solid Phase: Tags form duplexes with corresponding Antitags on the Microarray.



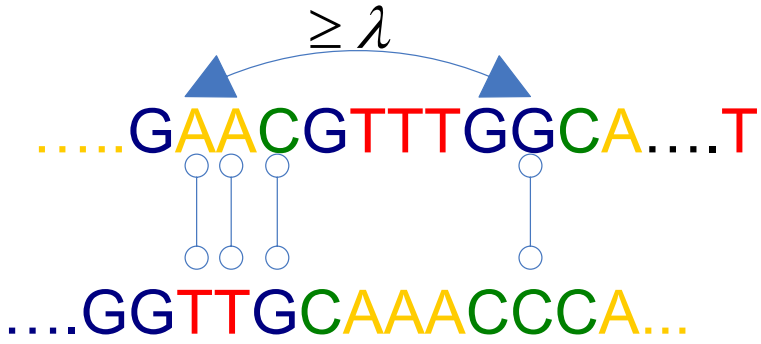
- Reporters are sorted into different locations on the array
- Dyes at each site determine the SNP variation of the tested sequence.

DNA TAT Design Tradeoff:

- Maximize the number of SNPs that can be genotyped in Parallel ≡Maximize the Number of tags
- Maximize the accuracy of the array ≡ Minimize Cross-Hybridization due to similar tags

Construct the Largest Possible λ Free Code over Σ={A,C,T,G}

- A stable duplex requires perfectly complementary Substrings of length more than λ



- λ-Free Code: No two elements of the set of tags of length n have a common substring of length more than λ.
- A De Bruijn sequence of order λ: A cyclic sequence in which each possible λ subsequence occurs exactly once.

- S. Brenner, “Methods for Sorting polynucleotides using oligonucleotide tags,” US Patent 5604097, 1997

- By parsing a De Bruijn Sequence of order λ: We get a λ-free code of size

$$4^{\lambda} / (n - \lambda + 1)$$

- The model does not take into account the thermodynamic properties of DNA duplexes.

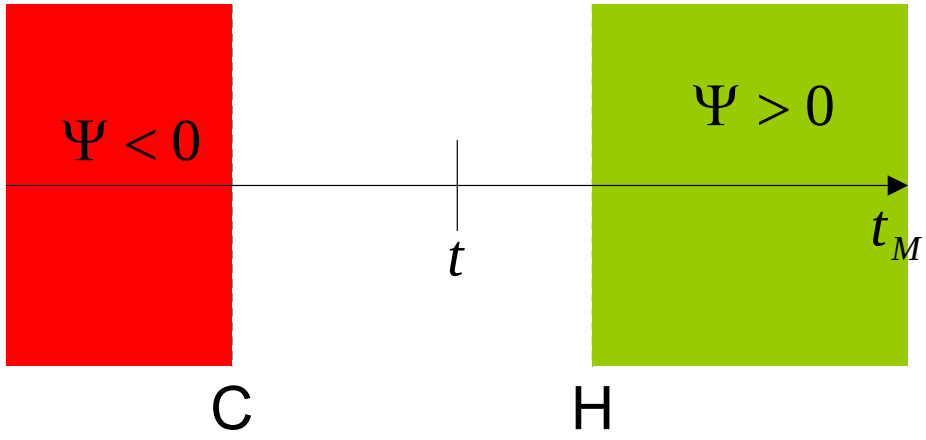
- Greedy (Heurestic) Algorithm:
  - Start with an empty set
  - Iteratively add a new tag provided that it does not hybridize with any of the tag complements already included
  - Allows the use of complex thermodynamic models
  - No analysis available.

M. Morris, D. Shoemaker, R. Davis and M. Mittman, “Methods and Components for selecting tag nucleic acids and probe arrays,” European Patent.

Thermodynamic Model

Melting temperaturebetween two sequences U & V: $t_M(U,V)$

- Half of the U and V oligoneucleotides will be in single stranded form and half occur in duplexes
- Measure of Hybridization, Ψ between U & V



- C-G pairs have one more hydrogen bond
- 2-4 Rule (for short oligonucleotides) [20]:
  $t_M(U,V)$=2(# of A-T pairs) + 4(# of C-G pairs)

TAT Design Problem:

- For each tag-antitag pair $t_M(U,\overline{U}) \geq H$
- For any two distinct tags *U & V*, and for each nucleotide substring *x*,
  $$x \in V \;\&\; x \in U, \quad t_M(x,\overline{x}) < C \;.$$
- Restriction [2]: x does not occur twice in the same tag if $t_M(x,\overline{x}) \geq C$ (not only in any 2 distinct tags).

- **Weight:** $w(s) = \sum_{i=1}^{k} w(s_i)$, **where** $s_i \in \Sigma = \{A, C, T, G\}$

$$w(C) = w(G) = 2w(A) = 2w(T) = 2.$$

Combinatorial Tag Design Problem:

- Find the maximum valid *c-h* code:
  - 1) Each tag has a weight *h* or more.
  - 2) Any substring of weight *c* or more occurs at most once.
    - Corresponds to *C=2c* and *H=2h*
  - 2') Any c-token occurs at most once

- *t* is a *c*-token: $w(t) \geq c$, No suffix of weight $\geq c$ is properly contained in *t*.
- Example: *c=4*

Weight	2 1 2 2 1 1 1	Tail Wt.
TAG	G A C C A A T	7
tokens:	C A A T	1
	C A A	1
	C C A	1
	C C	2
	G A C	2

- $\{A,T\} \to W, \quad \{C,G\} \to S$, $w(W) = 1, \quad w(S) = 2$
- $G_1 = 2, G_2 = 6, G_n = 2G_{n-1} + 2G_{n-2}, n \geq 3$

$$G_n = \left[\left(\frac{3 + \sqrt{3}}{6} \right) (1 + \sqrt{3})^n \right]$$

G2=6
A A
T T
A T
T A
C
G

- Lemma 1: Any tag in a valid c-h code has a tail weight of at least h-c+1
- Lemma 2: Total tail weight of a valid c-h code is at most $2G_{c-1} + 6G_{c-2} + 8G_{c-3}$

Token Class	Maximum Occurrence	Maximum Tail Weight	Token Weight
<c-2>S	$2\ G_{c-2}$	$4\ G_{c-2}$	c
S<c-3>S	$4\ G_{c-3}$	$8\ G_{c-3}$	c+1
<c-1>W	$2\ G_{c-1}$	$2\ G_{c-1}$	c
S<c-2>W	$2\ G_{c-2}$	$2\ G_{c-2}$	c+1

- Theorem 1: Any valid *c-h* code contains at most $\frac{2G_{c-1} + 6G_{c-2} + 8G_{c-3}}{h-c+1}$ tags.

- Optimal 4-10 code with 12 tags:

GACCAAT
CATTATCA
ATATAGTG
CAGCTAT
GAAATTCT
TAAAACTC
GTCGATA
CTTAATGA
AATAAGAG
CTGGTTA
GTATTTGT
TTTTACAC

- Construction Approach:
  - Construct a (maximal) set of circular strings over Σ where each c token occurs (at most) once
  - Extract minimal substrings of weight h or more with overlap at most c-1

- Greedy Algorithm
  - h=10, c=4

```
…CATTAATCAGCTATATAGTC….
   2  3  4 5  6  7 8 10
…C A T T A A T CAGCTATATAGTC….
..              4  3 2
…C A T T A A T CAGCTATATAGTC….
……………1  3  4 6  8  9 10
…CATTAAT C A G C T ATATAGTC..
…                    4  2 1
…CATTAAT C A G C T A TATAGTC….
..                     1 2  3  4 5  6  8  9 11
…CATTAATCAGCT A T A T A G T C….
```

  - Max tag weight =h+1
  - Min overlap weight= c-2
  - Max no overlap weight= h-c+3

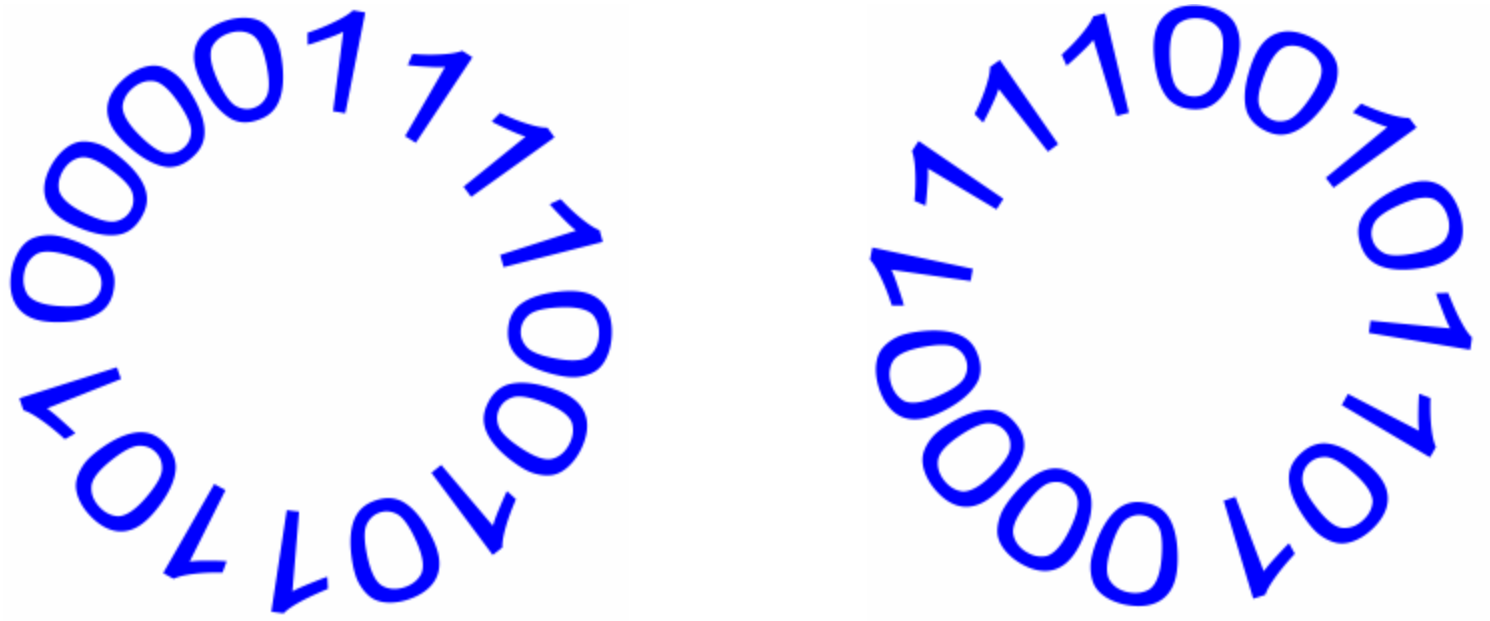
- Idea: Each metastring of weight c will be paired with a different bit string.

$$a = (\mu, \beta);\ a \in \Sigma,\ \mu \in \{W, S\},\ \beta \in \{0,1\}.$$

s:	A	T	C	G	(string)
μ :	W	W	S	S	(meta-string)
β :	0	1	0	1	(bit string)

- Idea: Use binary de-Bruijn sequences (D_k^i has length 2^k, each sequence of length k occurs once, and the linear string starts from offset *i*.

Rotating Drum Problem:



Inspired from `A Course in Combinatorics,' J. van Lint & R. Wilson

- Cycle Construction (Simple Case):
  - For metastring μ; $w(\mu) = c$(even):
    - $\gcd(|\mu| + 1, 2^{|\mu|}) = 1$
  - μw cannot be represented as a concatenation of two or more identical strings
- $C_0(\mu) = ((\mu w)^{2^{|\mu|}}, (D^0_{|\mu|})^{|\mu|+1})$
  - Example:

$\mu = SS, \quad |\mu| = 2, \quad D^0_{|\mu|} = 0011$

S S W S S W S S W S S W	$(\mu w)^{2\|\mu\|}$
0 0 1 1 0 0 1 1 0 0 1 1	$(D^0_{\|\mu\|})^{\|\mu\|+1}$
C C T G C A G G A C G T	$C_0(\mu)$
* * * *	

$\mu = SS, \quad |\mu| = 2, \quad D^0_{|\mu|} = 0011$

S S W S S W S S W S S W $(\mu w)^{2|\mu|}$

0 0 1 1 0 0 1 1 0 0 1 1 $(D^0_{|\mu|})^{|\mu|+1}$

C C T G C A G G A C G T $C_0(\mu)$

* ∧ * ∧ * ∧ * ∧

$\mu = SS, \quad |\mu| = 2, \quad D^0_{|\mu|} = 0011$

S S W S S W S S W S S W $(\mu w)^{2|\mu|}$

0 0 1 1 0 0 1 1 0 0 1 1 $(D^0_{|\mu|})^{|\mu|+1}$

C C T G C A G G A C G T $C_0(\mu)$

\+ * ∧ + * ∧ + * ∧ + * ∧

- *Observation:Total weight=Sum of tail weights.*

- **General Case:** $\mu w = (\alpha)^p, \quad k = \gcd\left(|\alpha|, 2^{|\mu|}\right)$

$$C_i(\mu) = ((\alpha)^{2^{|\mu|/k}}, (D^i_{|\mu|})^{|\alpha|/k}), \quad i = 0, \ldots, k-1$$

- **Example:** $c = 4, \mu = SWW, \alpha = SWWW, |\mu| = 3, k = 4$

$$D^0_3 = 10111000 = d_0 d_1 d_2 d_3 d_4 d_5 d_6 d_7$$

S	W	W	W	S	W	W	W
d_0	d_1	d_2	d_3	d_4	d_5	d_6	d_7
d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_0
d_2	d_3	d_4	d_5	d_6	d_7	d_0	d_1
d_3	d_4	d_5	d_6	d_7	d_0	d_1	d_2
		*			*		

S	W	W	W	S	W	W	W
d_0	d_1	d_2	d_3	d_4	d_5	d_6	d_7
d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_0
d_2	d_3	d_4	d_5	d_6	d_7	d_0	d_1
d_3	d_4	d_5	d_6	d_7	d_0	d_1	d_2
		+	*			+	*

S	W	W	W	S	W	W	W
d_0	d_1	d_2	d_3	d_4	d_5	d_6	d_7
d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_0
d_2	d_3	d_4	d_5	d_6	d_7	d_0	d_1
d_3	d_4	d_5	d_6	d_7	d_0	d_1	d_2
^	+	*		^	+	*	

S	W	W	W	S	W	W	W
d_0	d_1	d_2	d_3	d_4	d_5	d_6	d_7
d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_0
d_2	d_3	d_4	d_5	d_6	d_7	d_0	d_1
d_3	d_4	d_5	d_6	d_7	d_0	d_1	d_2
^	+	*	#	^	+	*	#

- c-tokens: SWW, WSW, WWS
- c+1-tokens: SWWW

- Recall:
  -

Token Class	Maximum Occurrence	Maximum Tail Weight	Token Weight
<c-2>S	$2\ G_{c-2}$	$4\,G_{c-2}$	c
S<c-3>S	$4\ G_{c-3}$	$8\ G_{c-3}$	c+1
<c-1>W	$2\,G_{c-1}$	$2\,G_{c-1}$	c
S<c-2>W	$2\,G_{c-2}$	$2\ G_{c-2}$	c+1

  - (Assuming) Total weight of Ω =sum of tail weights=
    $\Pi = 4G_{c-2} + 2G_{c-1} + 0.5\left(8G_{c-3} + 4G_{c-2}\right) = 2G_{c-1} + 6G_{c-2} + 4G_{c-3}.$
  - (Corrollary 2:) Total weight of Ω ,
    $$\Pi = 2G_{c-1} + 6G_{c-2} + 4G_{c-3}.$$
  - Max weight of extracted string with no overlap: *h-c+3.*
- *Theorem: The above construction yields at least*

$$\frac{\Pi}{h-c+3}$$ tags.

- Optimality of the construction:
  - In a valid set of cycles:
    - # of strong characters is upper bounded by G_{c-1}.
    - # of weak characters is upper bounded by G_c.
  - The total weight is upper bounded by G_c+2G_{c-1}.

Note that: $2G_{c-1}+6G_{c-2}+4G_{c-3}= G_c+2G_{c-1}$.

Proof:

- # of instances of $S^k W$ is (at most) $2^k G_{2r-2k}$
  - There are $2^{k+1} G_{2r-2k-1}$ tokens of weight *2r* of the type $< 2r-2k-1 > S^k W$
  - There are $2^{k+1} G_{2r-2k-2}$ tokens of weight *2r+1* of the type $S < 2r-2k-2 > S^k W$.
  - Each instance of $S^k W$ is the suffix of exactly one token.
- 2^r of the metastring S^r.
- # of strong characters is the sum of # of instances of S^r and of $S^k W$.
- # of weak characters is at most G_{2r} (k=0).
- $\sum_{k=1}^{r} 2^k G_{2(r-k)} = G_{2r-1}$
- The total weight is upper bounded by $G_c + 2G_{c-1}$.

Conclusions and Discussions

- Since the thermodynamic model is approximate violations may occur.
- Secondary structure in a tag may cause it to fail to hybridize to an anti-tag even though its weight> h
- Tag-foreign antitag pairs may occur due to near complementary sequences
- With the aid of computational and experimental approaches the violating tags may be deleted.