Probe design

for microarrays by H. Bjørn Nielsen

CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS

-What is a Probe

-Different Probe Types

-OligoWiz

-Probe Design

- -Cross Hybridization and Complexity
- -Affinity
- -Position

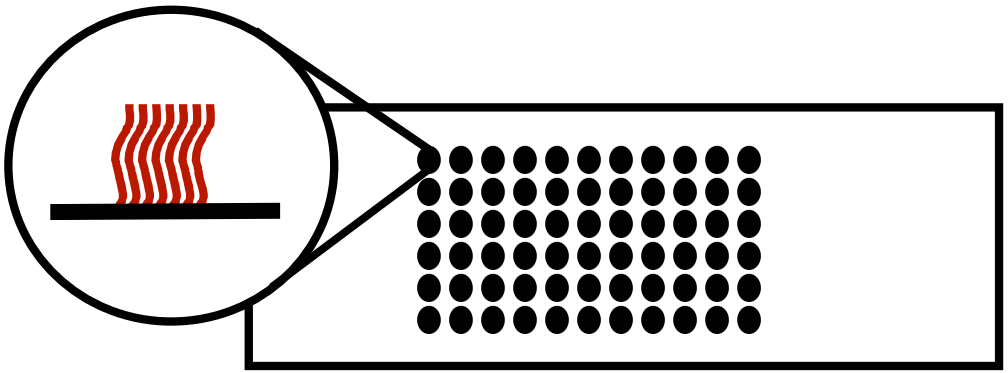
An Ideal Probe

must

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

- Discriminate well between its intended target and all other targets in the target pool

- Detect concentration differences under the hybridization conditions



Probe Type

comparisons

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

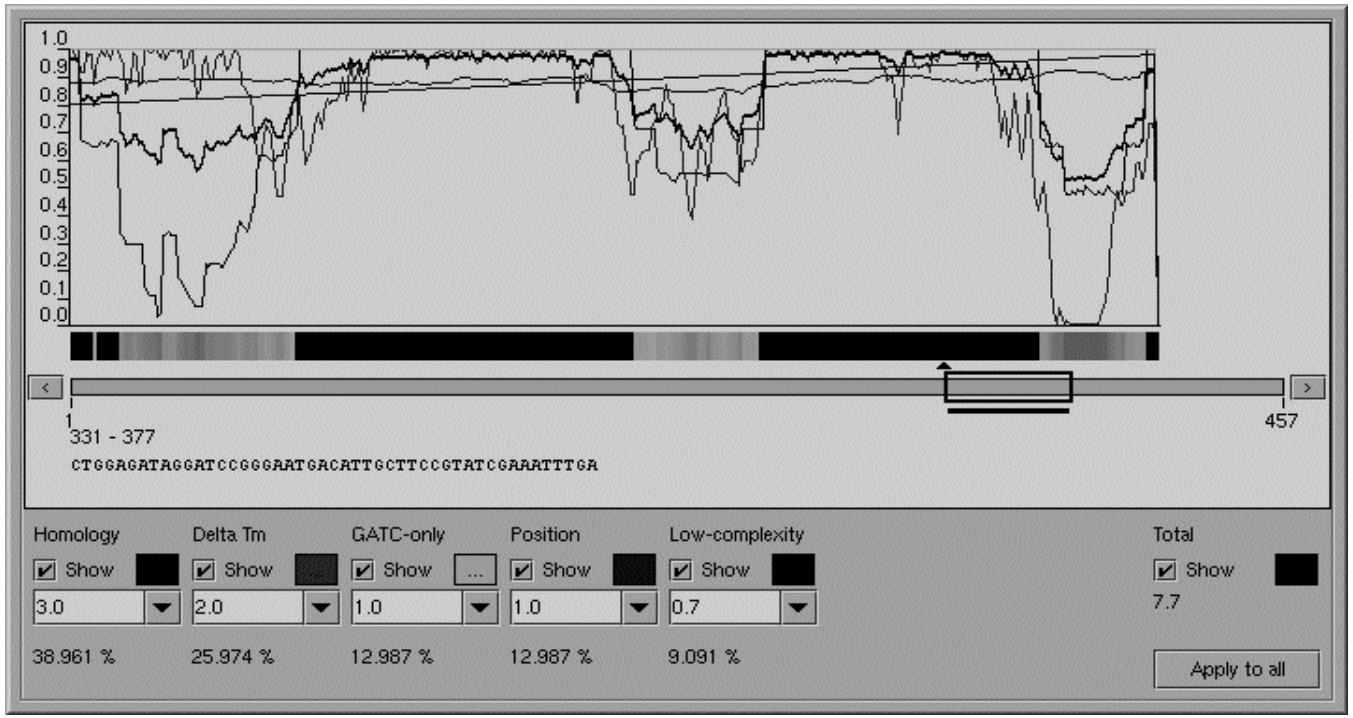
	Advantages	Disadvantages
PCR products	Inexpensive Linkers can be applied	Handling problems Hard to design to avoid cross-hybridization Unequal amplification
Oligos	Can be designed for many criteria Easy to handle Normalized concentrations Linkers can be applied	Expensive (Dkk. 100-150 per oligo)
Affymetrix GeneChip	High quality data Standardized arrays Fast to set up Multiple probes per gene	Expensive Arrays available for limited number of species

OligoWiz a Tool

for flexible probe design

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

A client-server application for designing oligonucleotides for microarrays



Download OligoWiz

and try the sample data set

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Go to

www.cbs.dtu.dk/services/OligoWiz/OW2

Linked from course program

Download the

- Program: OligoWiz2.jar
- Sample data set: yeast.fsa
- Annotation file: yeast.ann

Execute OligoWiz2 with the sample data set:

- Double click the icon (PC/Mac)
- java –jar OligoWiz2.jar (UNIX)

How to Avoid

cross-hybridization

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

From Kane et al. (2000) we learn that a 50'mer probe can detect significant false signal from a target that has

>75-80% homology to a 50'mer oligo

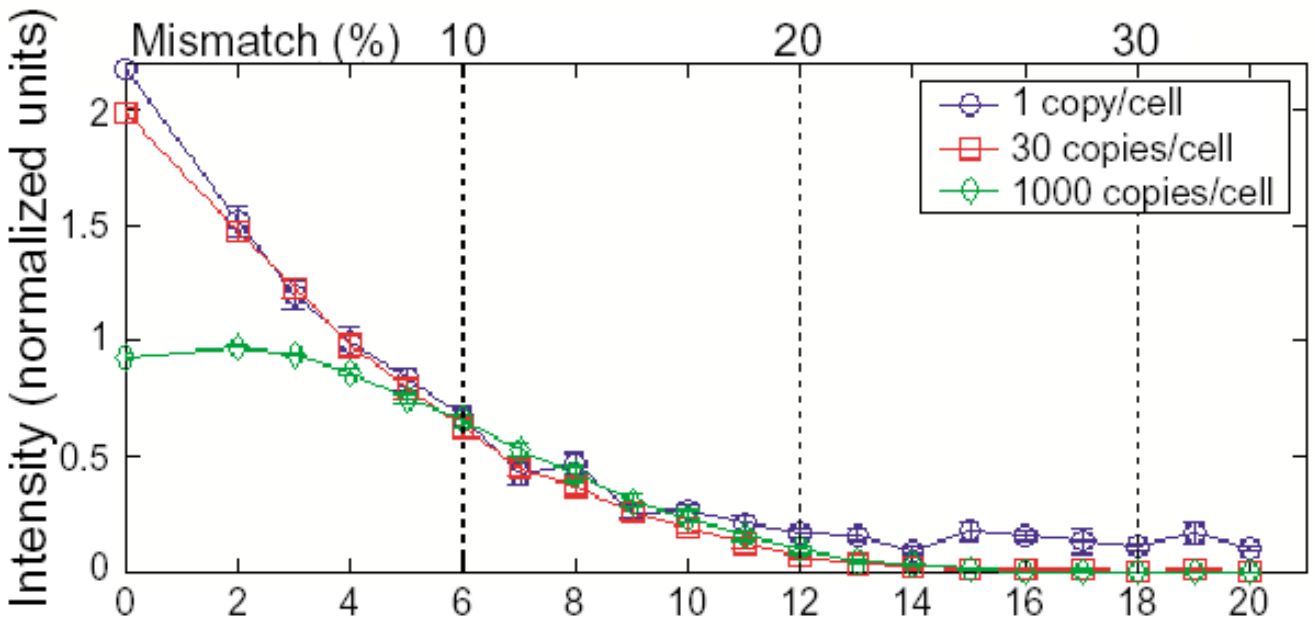
or a continuous stretch of >15 complementary bases

If we have substantial sequence information on the given organism, we can try to avoid this by choosing oligos that are not similar to any other expressed sequences.

Probe Specificity

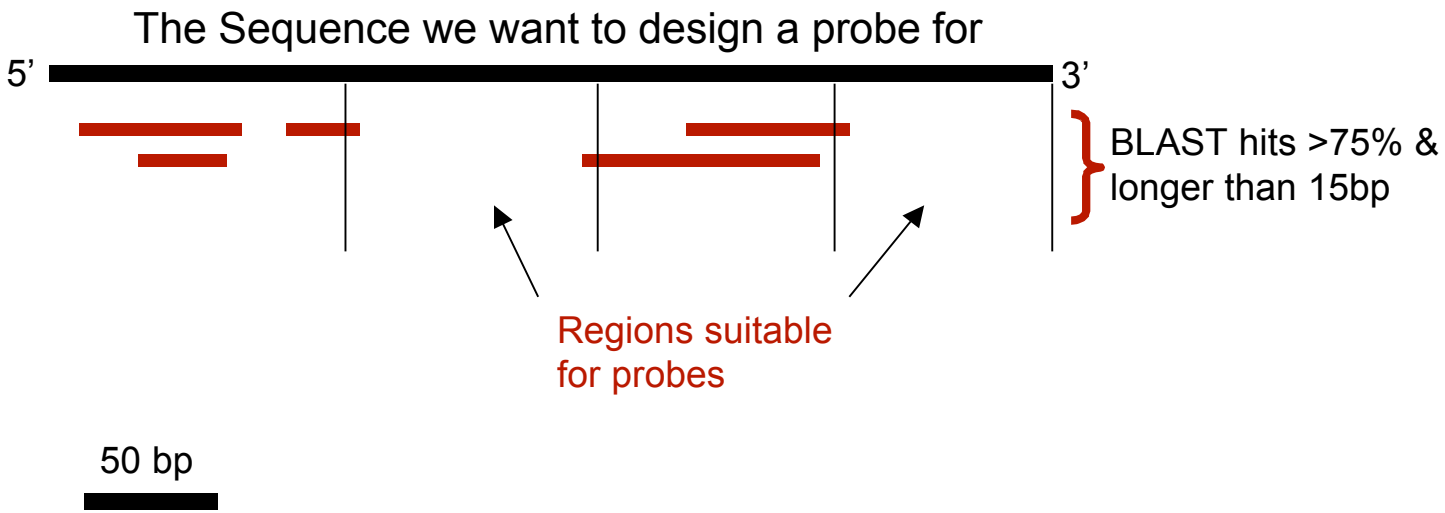
Hughes *et al.* 2001

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

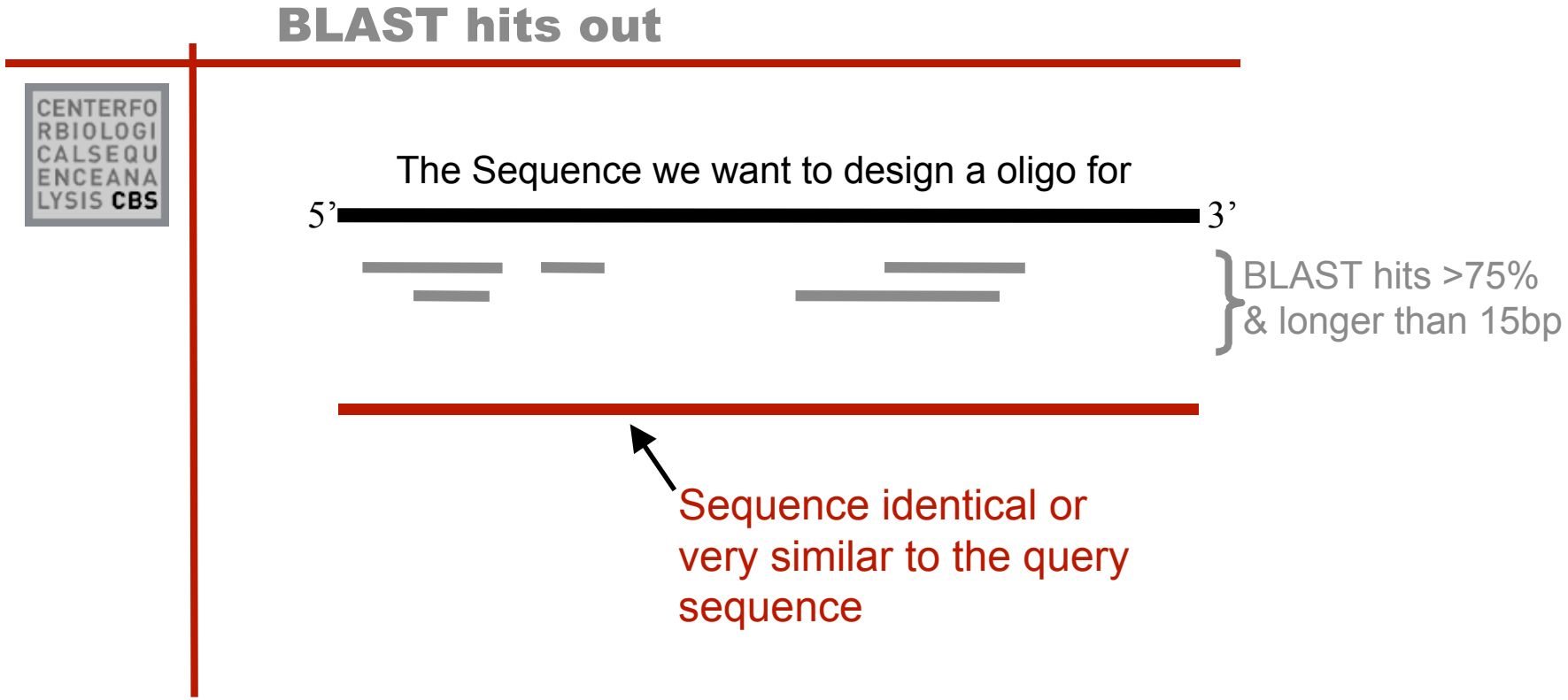


Mapping Regions

without similarity to other transcripts



Filtering Self Detecting



Therefore no BLAST hits with homology > 97% and with a 'hit length vs. query length' ratio > 0.8, are considered.

50 bp

Cross-hybridization

expressed as a 'homology score'

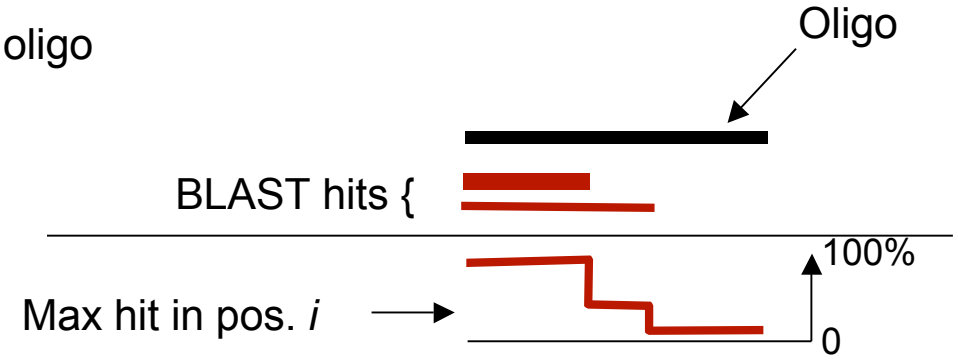
CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS

Only BLAST hits that passed filtering are considered

If m is the number of BLAST hits considered in position *i*.
Let h=(h1 $_i$,...,hm $_i$) be the BLAST hits in position *i* in the oligo

$$\text{BLAST max score} = \frac{100 \times n - \sum_{n}^{i=1} \max(h1_i, ..., hm_i)}{100 \times n}$$

Where n is the length of the oligo



Similar Affinity

for all oligos

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Another way of ensuring a optimal discrimination between target and non-target under hybridization is to design all the oligos on an array with similar affinity for their targets.

This will allow the experimentalist to optimize the hybridization conditions for all oligos by choosing the right hybridization temperature and salt concentration.

Commonly Melting Temperature (Tm) is used as a measure for DNA:DNA or RNA:DNA hybrid affinity.

Melting Temperature

difference

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

$$Tm(i) = \frac{1000\Delta H}{A + \Delta S + R\ln(\frac{Ct}{4})} - 273.15 + 16.6\log[Na+]$$

Where ΔH (Kcal/mol) is the sum of the nearest neighbor enthalpy, A is a constant for helix initiation corrections, ΔS is the sum of the nearest neighbor entropy changes, R is the Gas Constant (1.987 cal deg-1 mol-1) and Ct is the total molar concentration of strands.

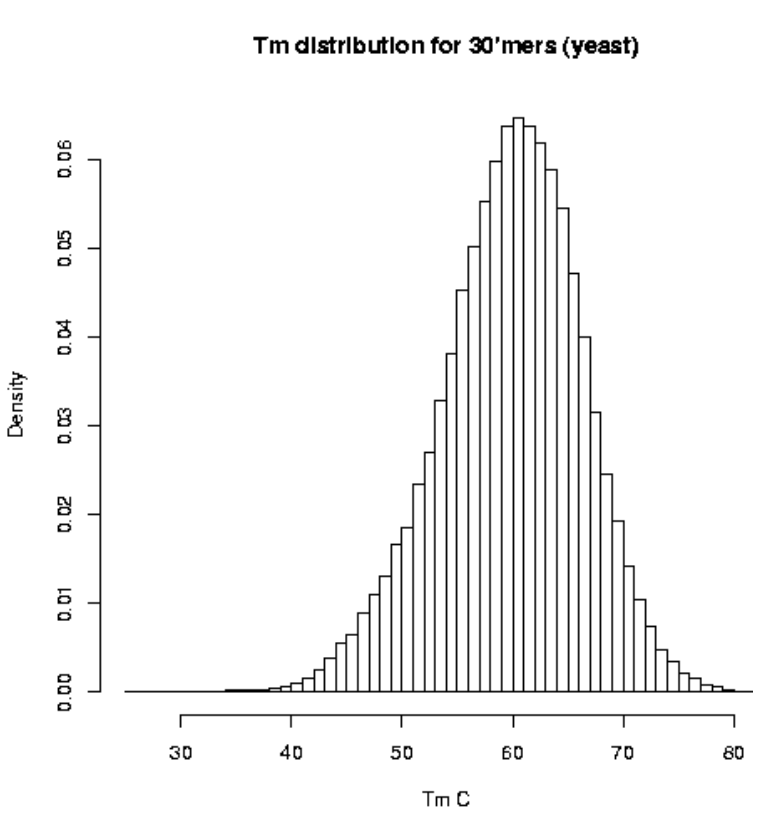
$$\Delta\mathrm{Tm\ score} = \left| \mathrm{Tm(i)} - \frac{\sum_{\mathrm{N}}^{1} \mathrm{Tm(i)}}{\mathrm{N}} \right|$$

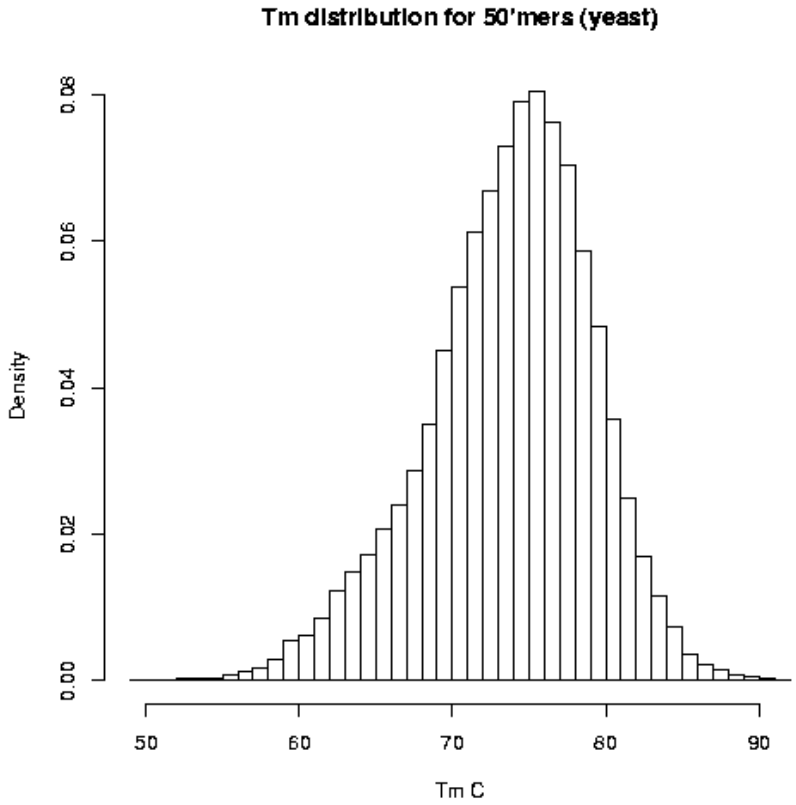
Where N is all oligos in all sequences.

Tm distributions

for 30'mers and 50'mers

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

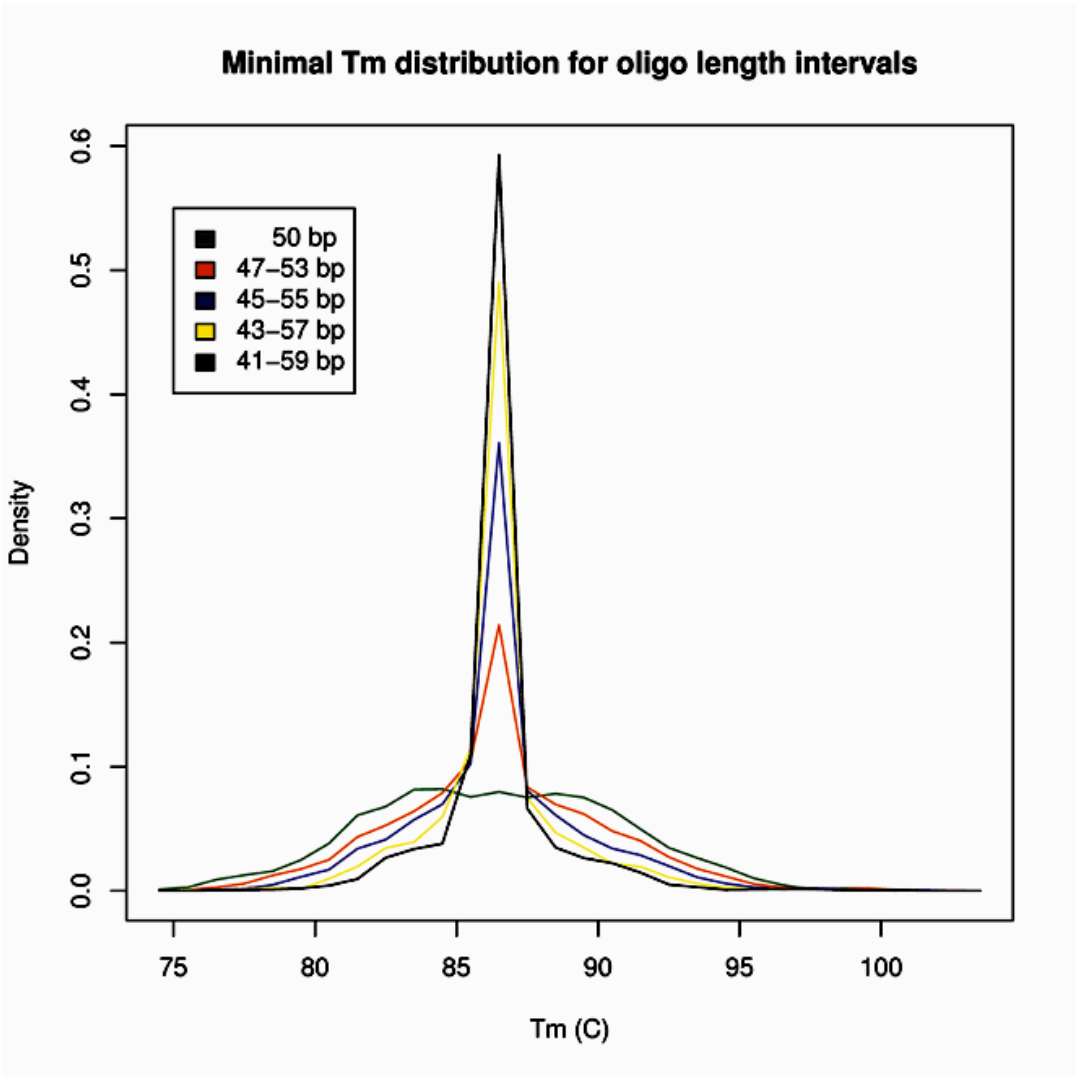




ΔTm Distribution

for oligo length intervals

CENTERFOR BIOLOGICAL SEQUENCE ANALYSIS CBS



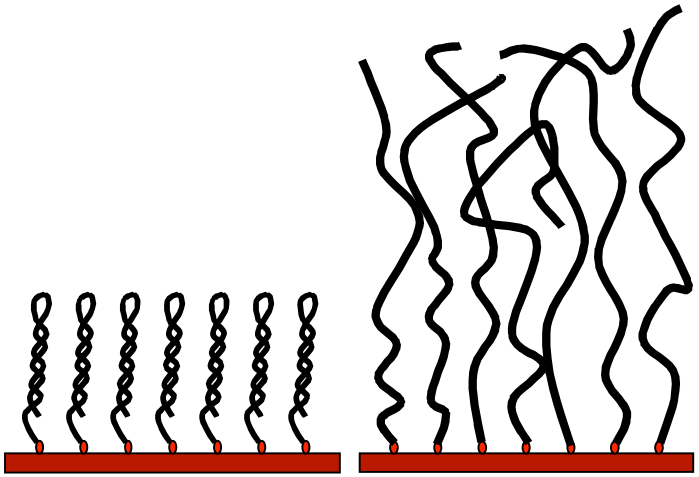
Avoid self annealing oligos

Sensitivity may be influenced

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Probes that form strong hybrids with it self *i.e.* probes that fold should be avoided.

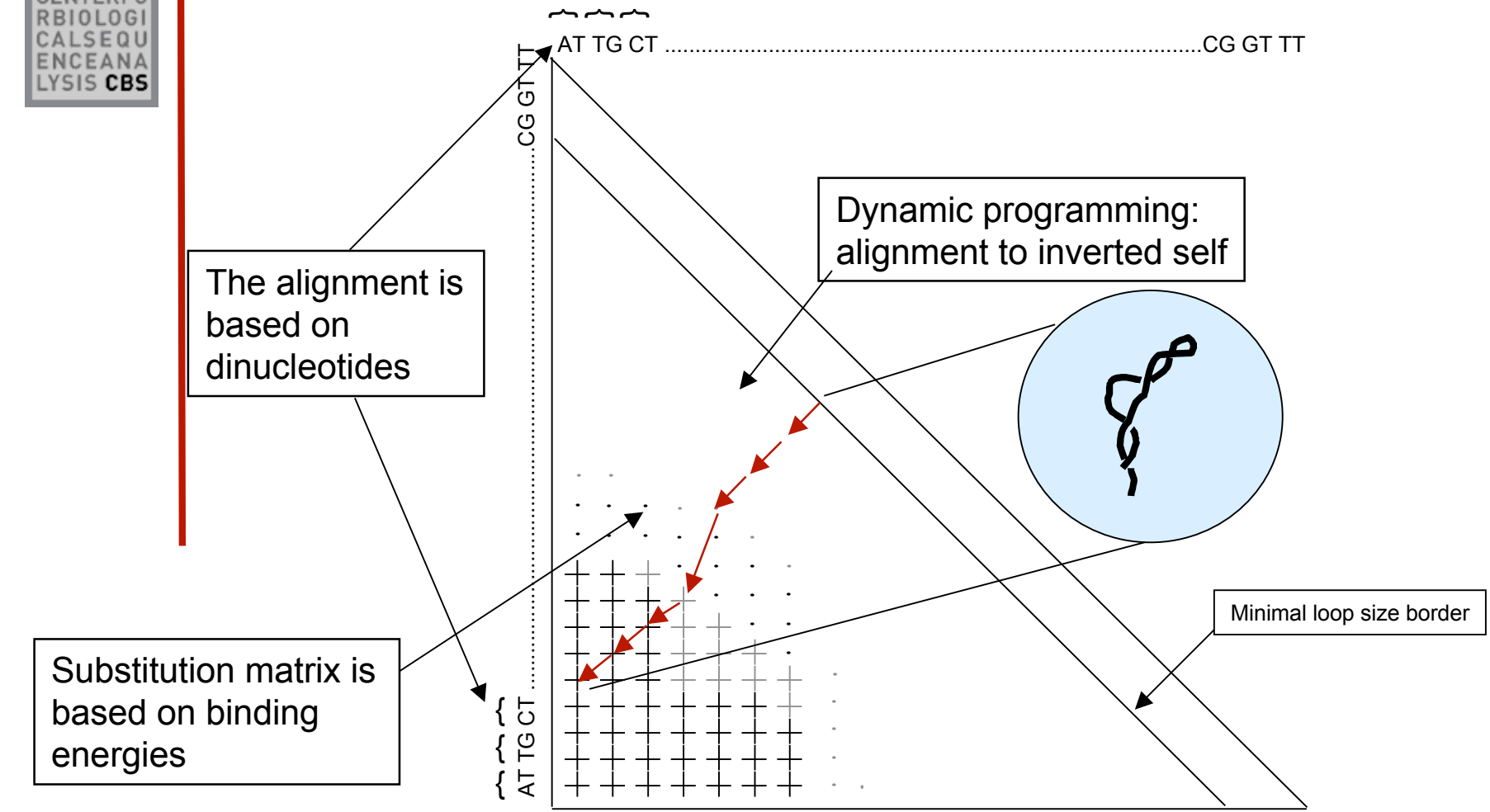
But, accurate folding algorithms like the one employed by mFOLD or RNAfold, is too time consuming, for large scale folding of oligos.



Time consumption:
mFOLD ~2 sec / 30'mer
Pr. gene (500bp) ~16 min.

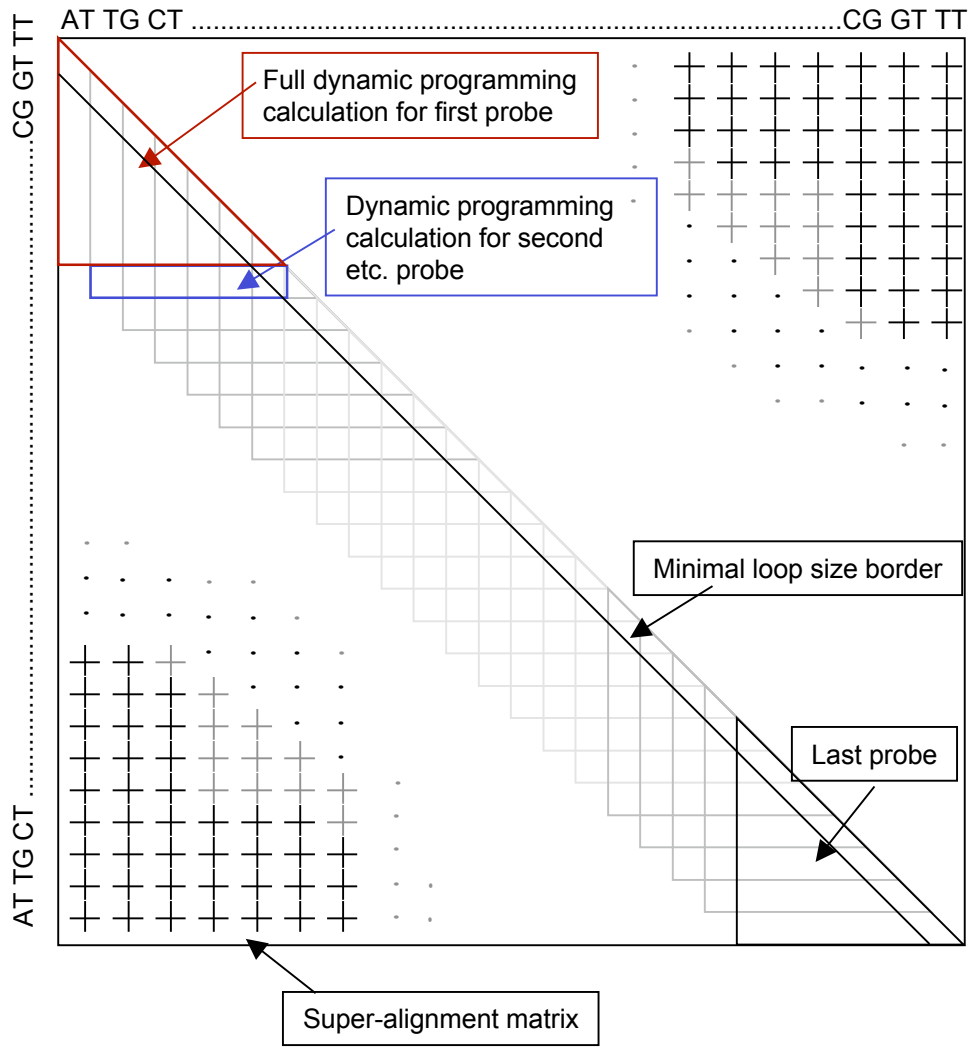
Folding an oligonucleotide

an approximation



Folding a lot of oligos

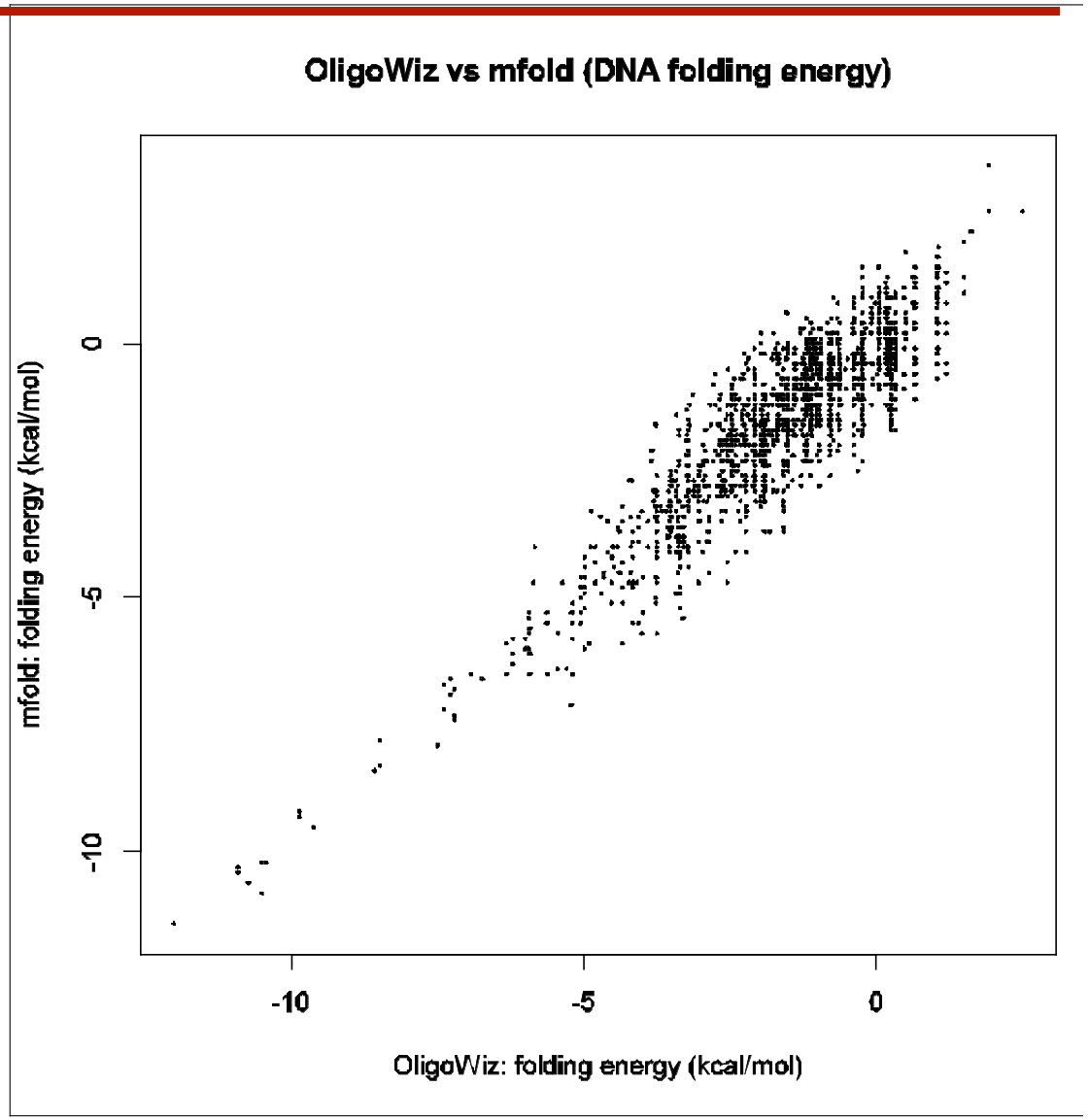
a fast heuristic implementation



Reasonably folding prediction

compared to mFOLD

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS



Probes With Very Common

sub sequences may result in unspecific signal

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

If the sub-fractions of an oligo are very common we define it as 'low-complex'

Oligo with low-complexity:
AAAAAAAGGAGTTTTTTTTCAAAAAACTTTTTAAAAAAGCTTTAGGTTTTTA
(Human)

Oligo without low-complexity:
CGTGACTGACAGCTGACTGCTAGCCATGCAACGTCATAGTACGATGACT
(Human)

Low-complexity

expressed as a score

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

For a given transcriptome a list of information content from all 'words' with length *wl* (8bp) is calculated:

$$I(w) = \frac{f(w)}{tf(w)} \log_2 \frac{f(w)}{tf(w)} 4^{wl}$$

Where *f(w)* is the number of occurrences of a pattern and *tf(w)* is the total number of patterns of length *wl*.

A low-complexity score for a given oligo is defined as:

$$\textbf{Low-complexity = 1-norm}\left(\sum_{L-wl+1}^{i=1} I(w_i) \right)$$

Where norm is a function that normalizes to between 1 and 0,
L is the length of the oligo and
W_i is the pattern in position i.

Location of Oligo

within transcript

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Labeling include reverse transcription of the mRNA and is sensitive to:

- RNA degradation
- Premature termination of cDNA synthesis
- Premature termination of cRNA transcription (IVT)

A ‘Position Score’ reflecting this (eukaryotes):

$$\text{Position score} = (1-drp)^{\Delta 3'end}$$

Where *drp* is the chance of labeling termination pr. base

Placement of multiple oligos

in respect to annotation

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Load annotation file:

Project/project properties : annotation file

Place 10 oligos in each gene:

Tools/oligo placement

Export the probes

to a file

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Export the oligos to a FASTA file:

File/Export oligo

Have a look at it

Place probes

relative to annotation

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Use the Tools/oligo placement tool, to place Oligos:

1) Exclusively in exons
2) Exclusively in introns
3) Exclusively on the exon-intron boundary

Using regular expressions

The “Introduction to regular expressions” Is linked from the course program.

Note: the yeast.ann file consists of lists of: {E,I,D,A,),(,. }
E=exon, I=intron, D= donor site, A=acceptor site,)(= end and beginning of exon and “.” = UTR