Useful truisms
A computer scientist looks at biology

JTS Aug 2005

(Some issues in biology, for the incoming computer scientist)

I. Generalities

II. An experimental/combinatorial procedure

Anticipating the development of biological knowledge

- **What we *need* to know**
- **What we *might be able* to know using presently imaginable techniques**
- ***Experimental techniques* for finding things out**

- ***The present and future biological databases***

A grossly oversimplified picture of the biological machine

- **A *program* is present on the DNA of each cell. Each gene of the program produces one protein (or a few variants), in a manner controlled by the binding of other proteins.**
- **The *program state* is represented by the concentrations and modifications of proteins and significant small molecules in each of the cell's compartments (nucleus, cytosol, cell membrane, other organelles.) With some state represented by epigenetic modifications, e.g. age may be represented in DNA.**
- **'Flag' proteins set up stable, 'determined' cell states.**
- **The cell moves between steady states in response to external stimuli, and also has one universal dynamic cycle, the mitotic (cell division) cycle. (With a bit more dynamics e.g. development, circadian clock.)**

What we need to know (at a minimum)

- **Hex dump of the program (the genome and its genes - available; 6,300 for yeast)**
- **Modification list per protein (components of state)**
- **List of significant small molecules, modification list per molecule**
- **Interaction matrix (M_{ij}; i,j index all proteins, modified proteins, and small molecules)**
- **Interaction outcomes matrix ([i',j'] = O_{ij})**
- **Modified state half-lives, interacting protein pair half-lives (H^*_i,H_{ij})**
- **Cell membrane receptors list**
- **Response states list (to receptors turned on; receptor interactions)**
- **Mitiotic cycle movies (states during each phase of cycle)**
- **Secretion rates**
- **Pathways (internal linkages; illuminated by interaction outcomes matrix)**
- **The internal 'robotics' of the cytoskeleton**

What we might be able to know soon

- **Complete genome (high-throughput methods available)**
- **Complete transcriptosome (high-throughput methods available)**
- **Complete physical collection of 'raw' proteins (gene expression in bacteria)**
- **Complete interaction matrix for raw proteins (high-throughput methods available)**
- **Extensive interaction outcomes matrix**
- **Part of modification list per protein (components of state)**
- **List of significant small molecules, partial modification list per molecule**
- ***But: not clear how to find high-throughput techniques for protein modifications***
  - **Significant protein modifications: phosphorylation, methylation, acetylation, ubiquitination (at least 200 of these modifications are known)**
  - **Glycosylation introduces trees of highly variable structure as covalently bound modifiers.**
  - **Post-translational protein modifications are invisible in genome.**
  - **Not clear how modified proteins can be produced in high-throughput fashion.**

Biological knowledge needs to be systematic

- **Episodic studies**
  - **warn us of what we may need to allow for in systematic accounts**
  - **are just 'snapshots' of situations of interest**
    - **'keyhole view'**
  - **But: may have special medical or other application interest**
  - **But: may yield crucial experimental techniques (e.g. enzymes)**

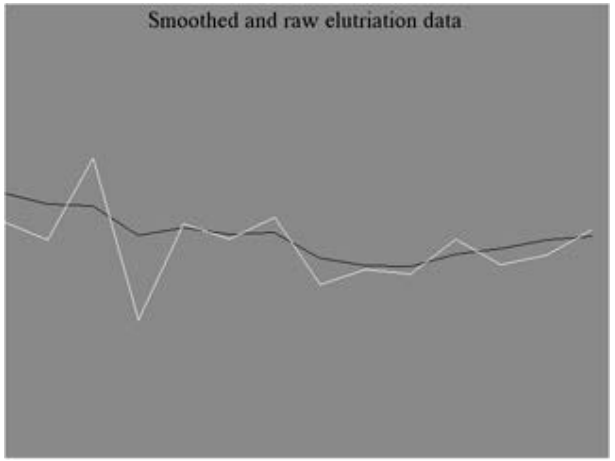
- **Systematic accounts**
  - **yield perspective**
  - **tell us what we do not need to to allow for**
  - **provide 'maps' and 'blueprints'**
    - **Broad and steady view**

- **The software tools used need to move from a focus on retrieval of individual individual facts to tools for survey of extensive 'fact landscapes'**

  - **Provided by high-throughput (often parallel) experimental techniques**

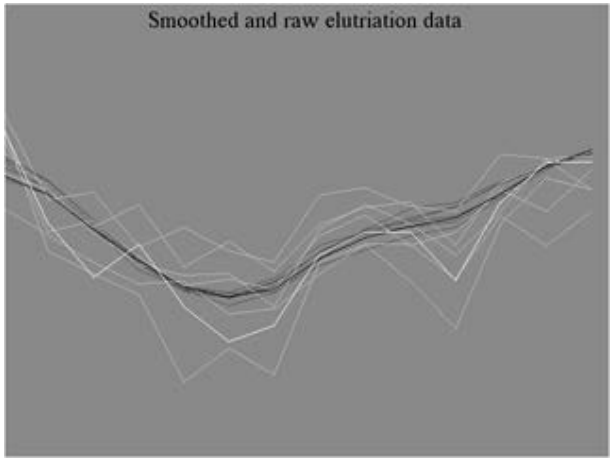
A small example: gene expression data for ~6,300 yeast genes

- **From Michael Eisen's lab at LBL**
- **6,300 by 80 matrix of measurements, divided into 6 separate 'time series'**
  - **Cell cycle (1 - 4), sporulation, diauxic shift**
- **The data is noisy, but as a time series can be smoothed in obvious ways**



center the average value, normalize the total above-average sum, and take a 5-point weighted moving average

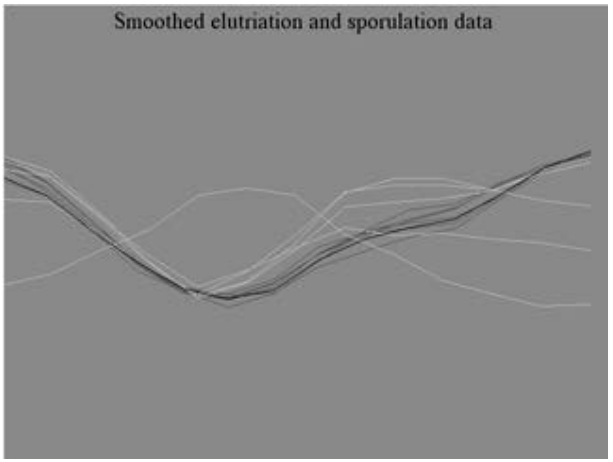
Clustering depends critically on data preparation



Tools for rapid data inspection may be preferable to automated data-analysis tools

In this example

- **Persuasive clusters appear in the smoothed data**
- **Different, and less persuasive clusters appear in the raw data**
- **The clusters seen in a given time series, which argue for a common control mechanism, are not always confirmed in other series**
- **Data for cross-species comparison of these expression profiles needs to be collected**



- dKlPaltYHcsVELnGniYIfGGLmPcYsYEEDAPMLndFfVDGIKNlPPPLLpQvINNP Cerevisiae
- yKpPsflYHtaVELaGniYIlGGLiPiYgYEEDAPDLsqFkVDGIKNiPPPLLpQiINNP Glabrata

The computer scientists' role

- **Help organize the masses of data which will accumulate**
  - **Devise effective techniques for search, inspection, comparison**
    - **e.g. BLAST**
    - **Interactive graphical interfaces**
  - **Devise computationally effective means for exploitation of available data *(well represented by Ron Shamir's talk at last year's conference)***
- **Simulation tools for sources of experimental error and for theories**
- **Help design and optimize high thoughput experimental approaches, which sometimes share the combinatorial flavor of parallel algorithm design.**
  - **'the technique is clever, but is it best possible?'**
- **Process automation is an important issue: robotic system and software design.**

Remainder of the talk

- **An ingenious *high throughput* procedure (MPSS)**
  - **Which biologists will recognize as a familiar but elaborate kind of high-throughput experimental DNA analysis**
  - **And which computer scientists will recognize as a parallel string-processing algorithm (which they may be able to analyze and optimize).**
  - **The intellectual skills involved in high-throughput experiment design and in parallel algorithm design are analogous.**

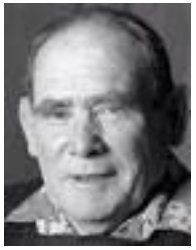
MPSS: massively parallel signature sequencing

(in a microarray variant)

Counts RNA (converted to cDNA) molecules in a population of ~10,000 known species, present in varying numbers representing 'expression levels', to a total of 1,000,00 items.

Tools:
- ***Microarrays*, allowing highly parallel (x1,000,000) *observation***
- ***Enzymes*, for *manipulation* of RNA and DNA strings.**

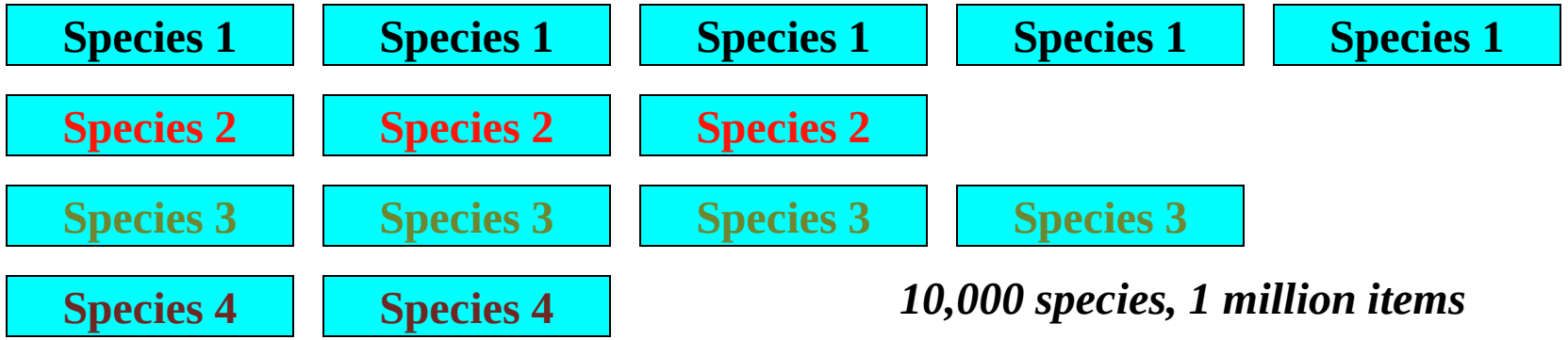
1) **Customize a microarray chip with 1M spots, each containing a distinct 10-base tag.**
2) **Hybridize these same tags, in random, blunt-end fashion, to a population P of cDNA molecules using which the expression levels of a smaller number of RNA species are to be measured.**
3) **Randomly select 1 million of the hybrids and amplify them with PCR, thereby producing a random collection S of the hybrids, each represented in numerous identical copies.** ***(For details, see later slide.)***
4) **Hybridize these to the chip; each spot will collect only those hybrids which have ends which match the spot.** ***(For details, see later slide.)***



Home

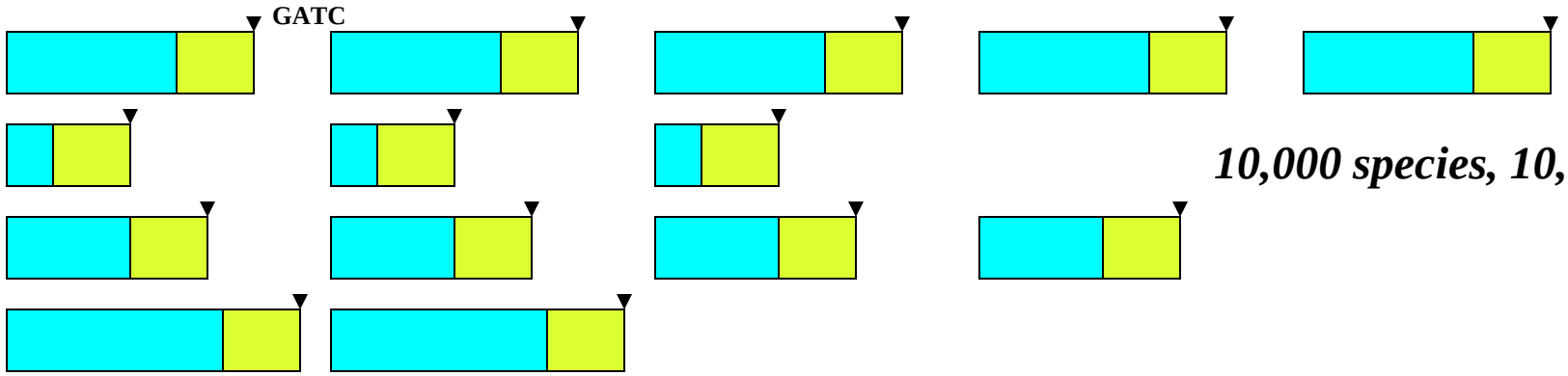
NATURE BIOTECHNOLOGY VOL 18 JUNE 2000

Details (1): The population of molecules to be counted



10,000 species, 1 million items

Details (2): The 'id' subsections identifying these molecular species



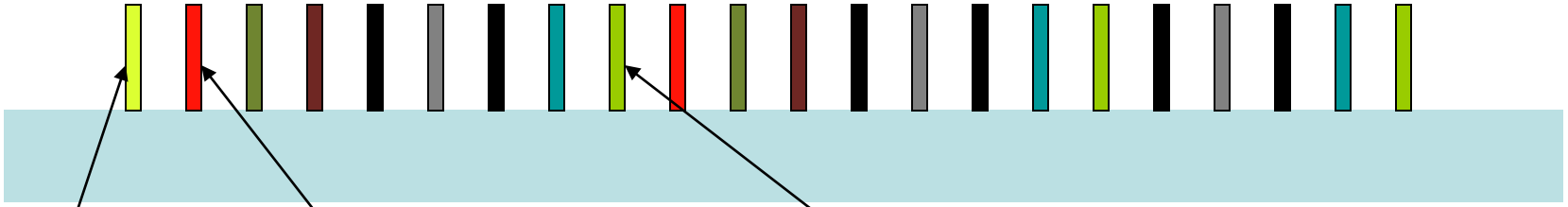
10,000 species, 10,000 ids

'ids' are sections of fixed length just to the left of fixed enzyme cut site at occurrence of known 4-letter pattern

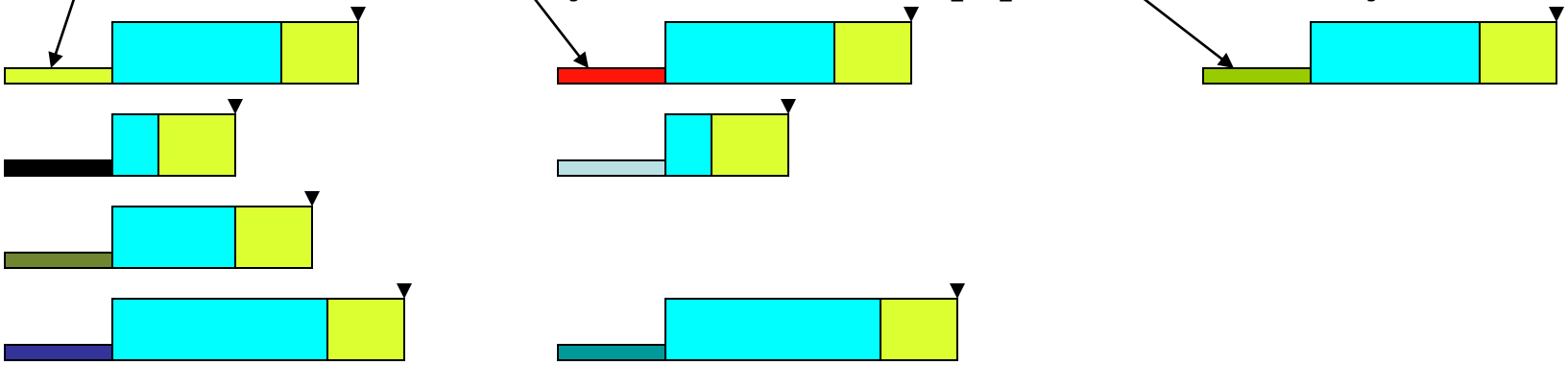
12 bases read in the yellow sections will identify species in a population of 10,000 species

Details (3): Technique for reading the tag sections

Using a customizable microarray of 1 million spots, all different

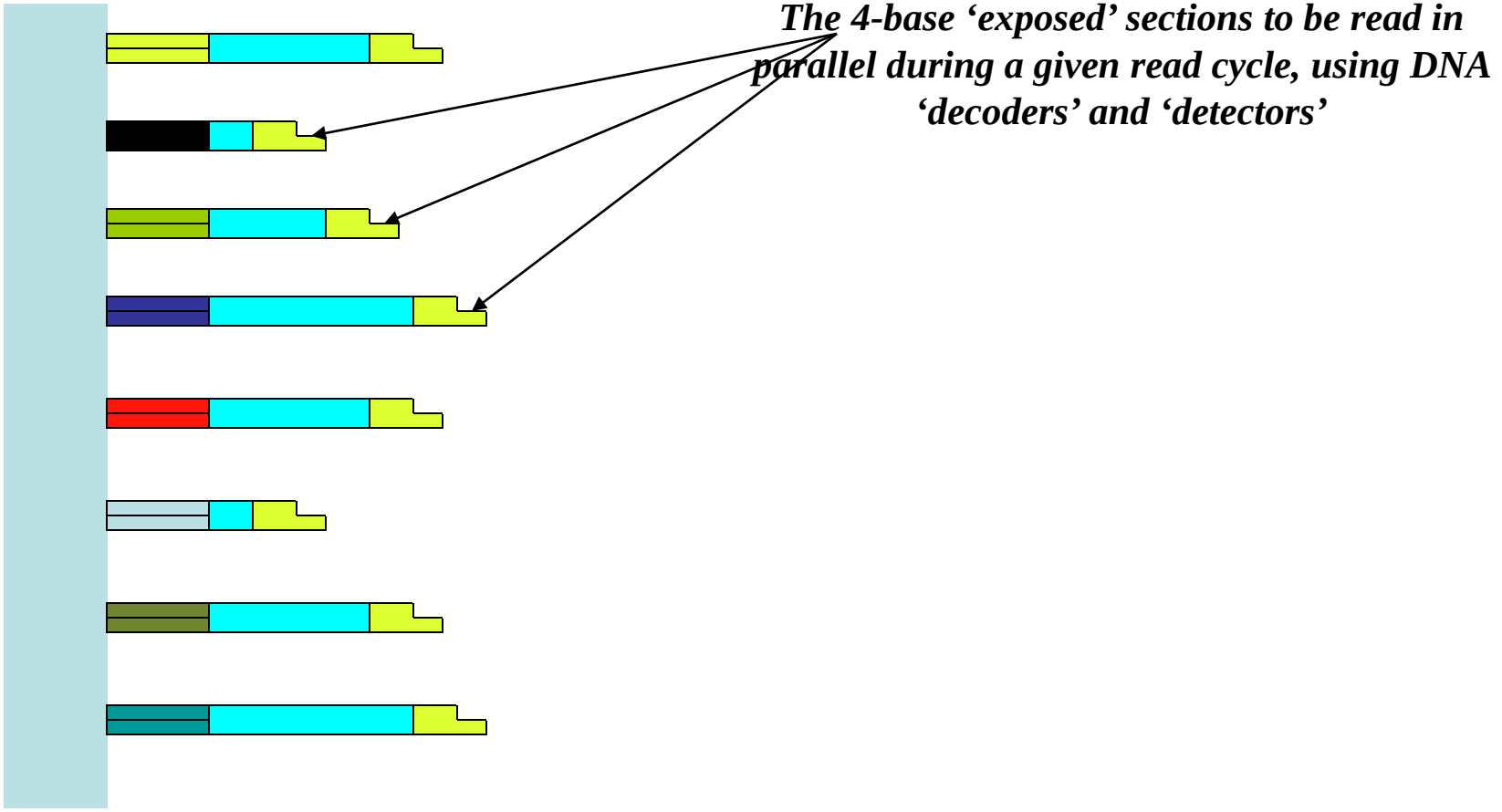


Details (4): **Amplify the collection of complements of these spotted sequences, and attach them randomly to the ends of the population to be analyzed.**

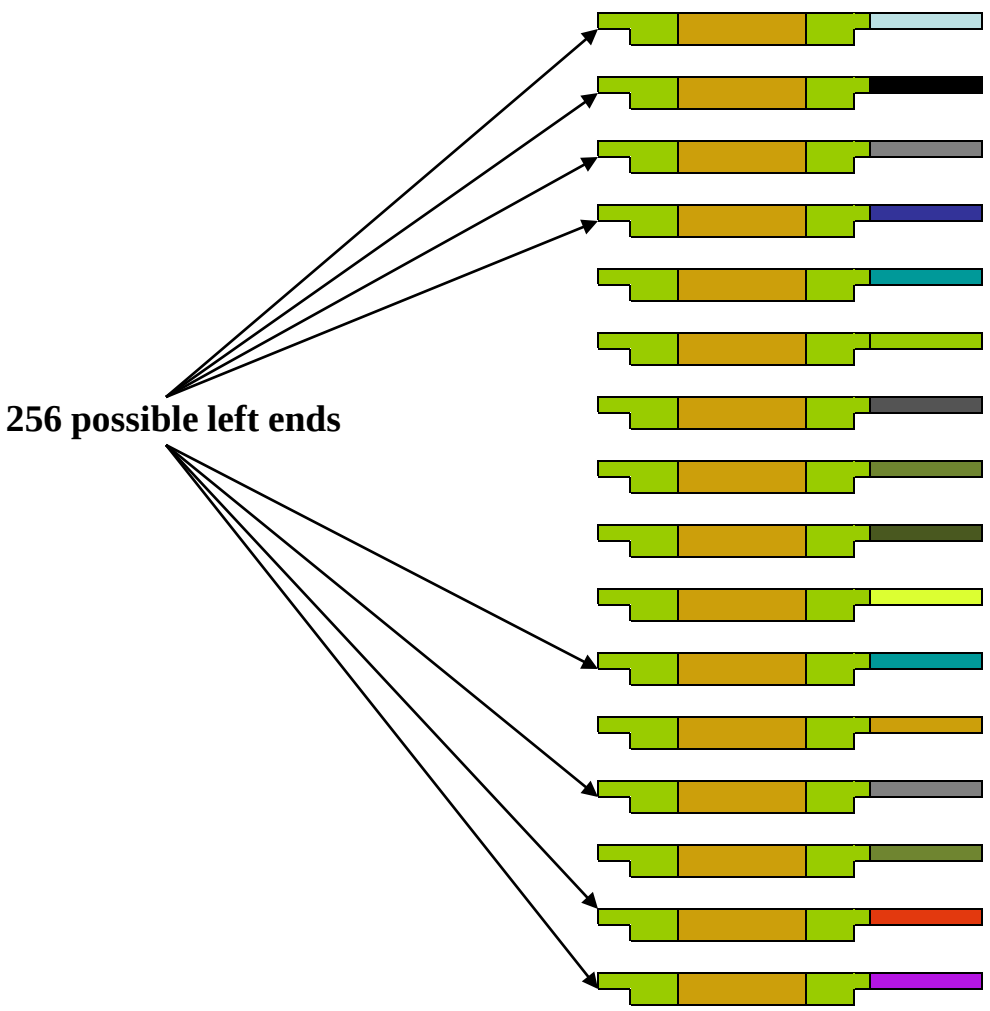


This forms a collection of 1,000,000 • n_molecular_species of molecules, of which 1,000,000 are selected (by dilution and re-amplification).

Details (5): After cutting 4 (unknown) bases into the population to be read, hybridize them to the microarray



Details (6): Combinatorial design of the 1024 'decoders': 16 right ends, each attached to 256 left ends



Right end [1,1]. ***Attaches to all left ends with $B_1 = 1$***

Right end [1,2]. ***Attaches to all left ends with $B_1 = 2$***

Right end [1,3]. ***Attaches to all left ends with $B_1 = 3$***

Right end [1,4]. ***Attaches to all left ends with $B_1 = 4$***

Right end [2,1]. ***Attaches to all left ends with $B_2 = 1$***

Right end [2,2]. ***Attaches to all left ends with $B_2 = 2$***

Right end [2,3]. ***Attaches to all left ends with $B_2 = 3$***

Right end [2,4]. ***Attaches to all left ends with $B_2 = 4$***

Right end [3,1]. ***Attaches to all left ends with $B_3 = 1$***

Right end [3,2]. ***Attaches to all left ends with $B_3 = 2$***

Right end [3,3]. ***Attaches to all left ends with $B_3 = 3$***

Right end [3,4]. ***Attaches to all left ends with $B_3 = 4$***

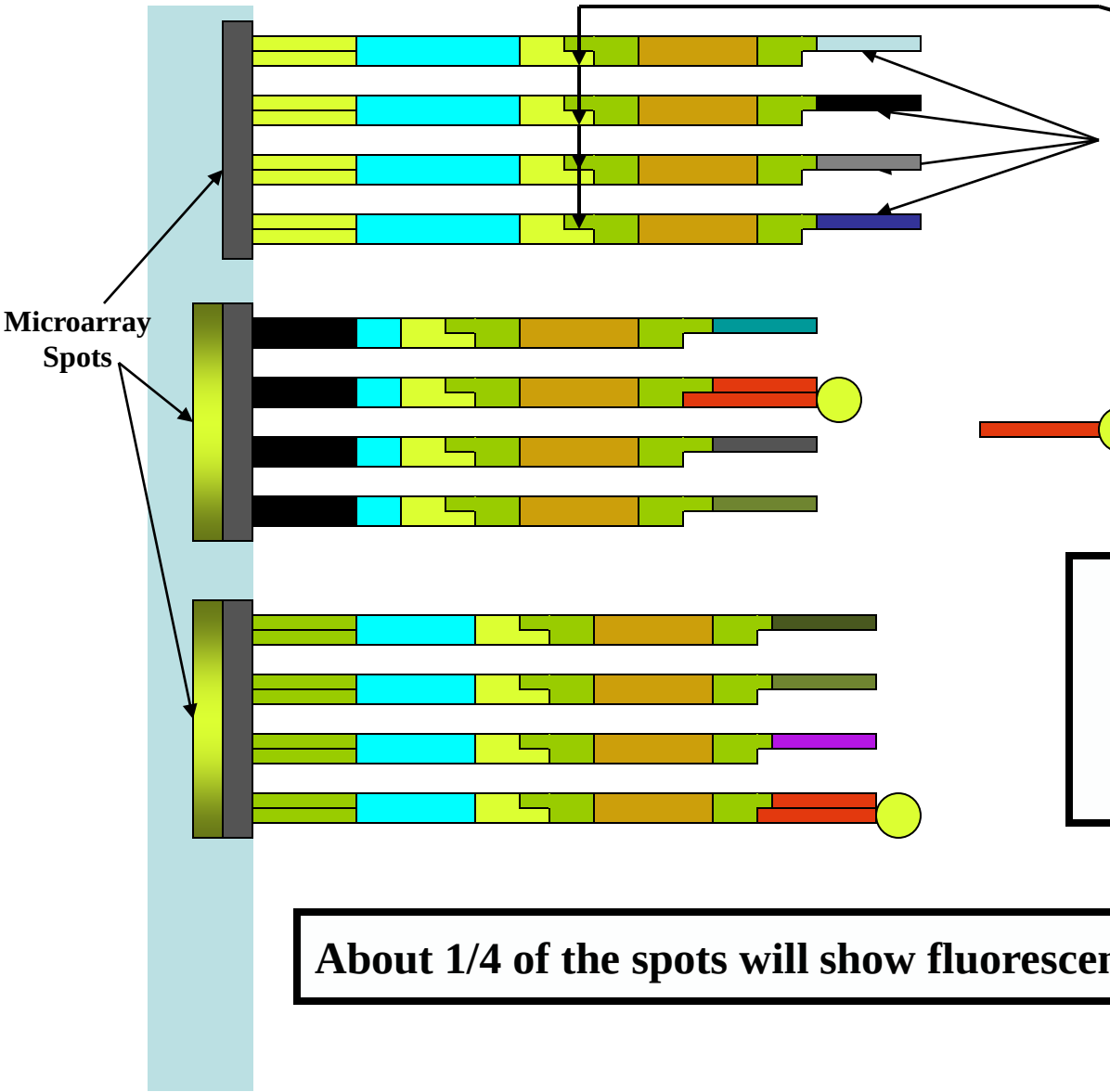
Right end [4,1]. ***Attaches to all left ends with $B_4 = 1$***

Right end [4,2]. ***Attaches to all left ends with $B_4 = 2$***

Right end [4,3]. ***Attaches to all left ends with $B_4 = 3$***

Right end [4,4]. ***Attaches to all left ends with $B_4 = 4$***

Details (7): **Exactly 4 of the 1024 'decoders' will hybridize to each microchip spot, causing exactly 4 of the 'detectors' to attach subsequently.**



The left-end pattern $B_1B_2B_3B_4$ determines which 4 right-end patterns $P_{j,Bj}$ j = 1..4 occur.

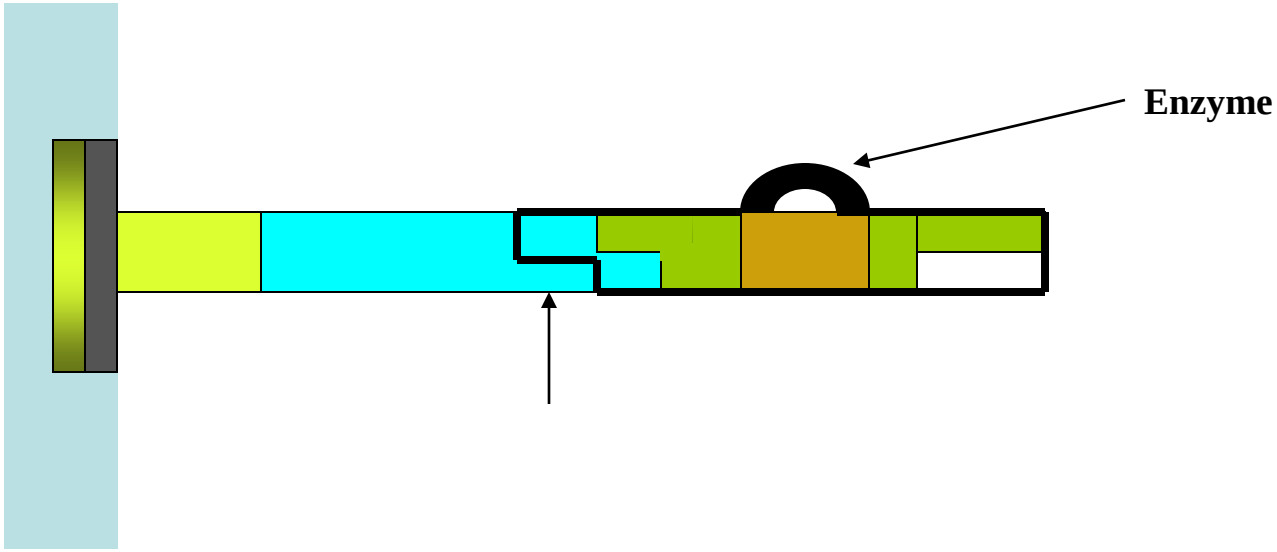
The 16 'detectors' are *complementary* to the 16 right-end patterns $P_{j,Bj}$, and *fluorescently* tagged.

The 16 'detectors' are hybridized and observed via fluorescence, one after another, and then washed off

About 1/4 of the spots will show fluorescence during each observation

Details (8): Moving on to the next 4 bases.

The (otherwise unused) central section of each decoder must be a recognition site for an off-center cutting enzyme that will cut to the left of the center, leaving the next 4 adjacent bases exposed for reading by reattachment of a new collection of decoders.



This 4-bases-at a time read cycle can be repeated as often as needed.

Analysis of combinatorial efficiency

- **The [1,000,000•10,000] Cartesian product of [tags•population] is formed, and 1,000,000 random elements of this collection are then selected for processing.**
- **A given tag may not be selected at all, or may be selected just once, twice, etc. If it is selected just a few times, the attached population elements are all distinct with high probability. The probability that a tag is selected n times is $e^{-1}/n!$ by the law of small numbers.**
- **We can tell how many population elements have become associated with a given tag by whether the bases in read in the corresponding position are unique, 2-ways, 3-ways, or 4-ways ambiguous. These imply unique, double, triple, or more than 4-fold associations with the tag for the position.**
- **Thus e^{-1} = 37% of the spots on the microchip carry no population element, 37%carry just one, 14% carry just two, and the remainder carry 3 or more.**
- **If only the unambiguously populated spots are read our we get 370,000 reads per experiment, but using the dually populated spots raises this to 740,000.**

The End

Even where the basic mechanisms involved are known or surmised, there may lie difficult-to explore continents of fact

- **E.g. implications of the 'protein modifications problem'**

- **We want to know all the components of the (huge, sparse) interaction matrix M_{ij} , but don't even know all the indices i and j.**

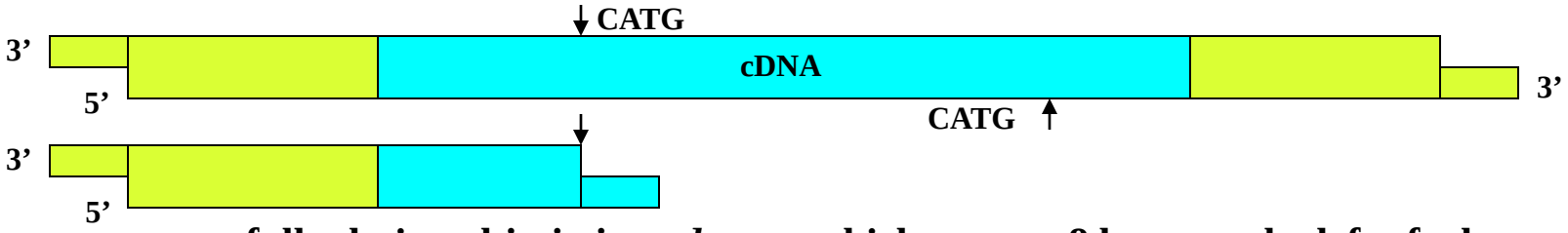
- **Only the first 50,000 of 50000•n interaction components lie even presumptively in reach of current high-throughput techniques.**

MPSS: details of the starting steps

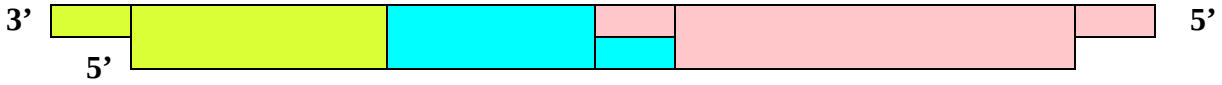
1) **Collect the cDNA population P and the collection T of 1 million distinct tags.**

2) **Cut the tags at one end to create an overhang, then blunt-end ligate P to T.**

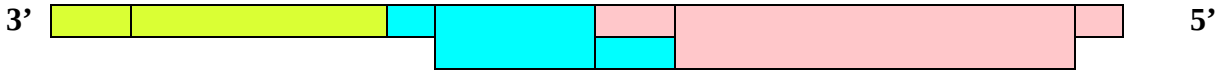
3) **Cut all the cDNAs in P at every CATG occurrence, leaving a 4-base underhang.**



4) **Ligate on a carefully designed *intiating adaptor*, which can cut 8 bases to the left of where it is attached; cut at a spacer appended to it, to leave a 5’ (immunizing) overhang.**



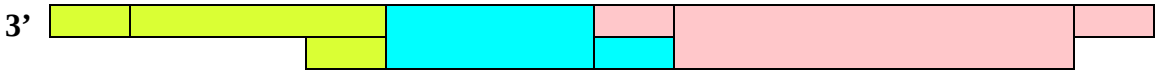
5) **Peel back the 5’ end at the 3’ overhang to expose the whole of the tag.**



T7 Exonuclease

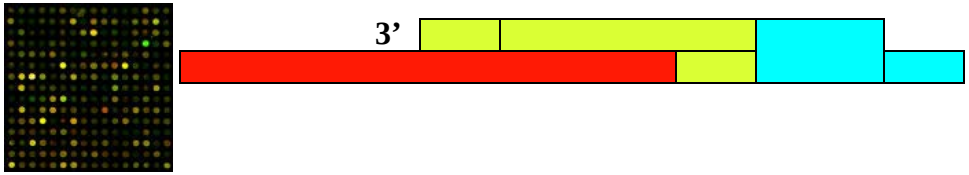
MPSS: details of the starting steps, contd.

- **Rebuild if necessary to have the whole cDNA element. Save this ‘soup’.**



- **Collect a *sample* of 1 million tagged cDNAs, and amplify by PCR. Use the initiating adaptor to cut 8 bases into the cDNA, leaving a 4 base underhang.**

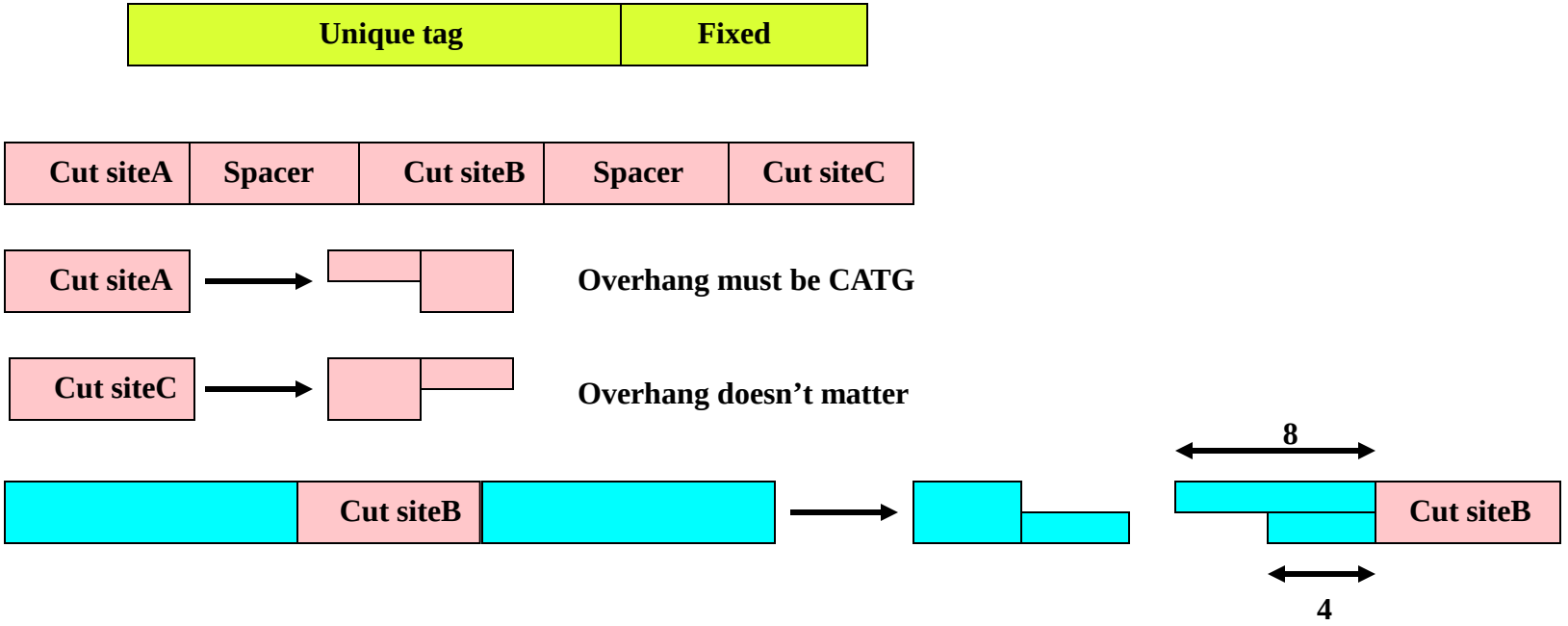
- **Hybridize to the chip, and begin reading.**



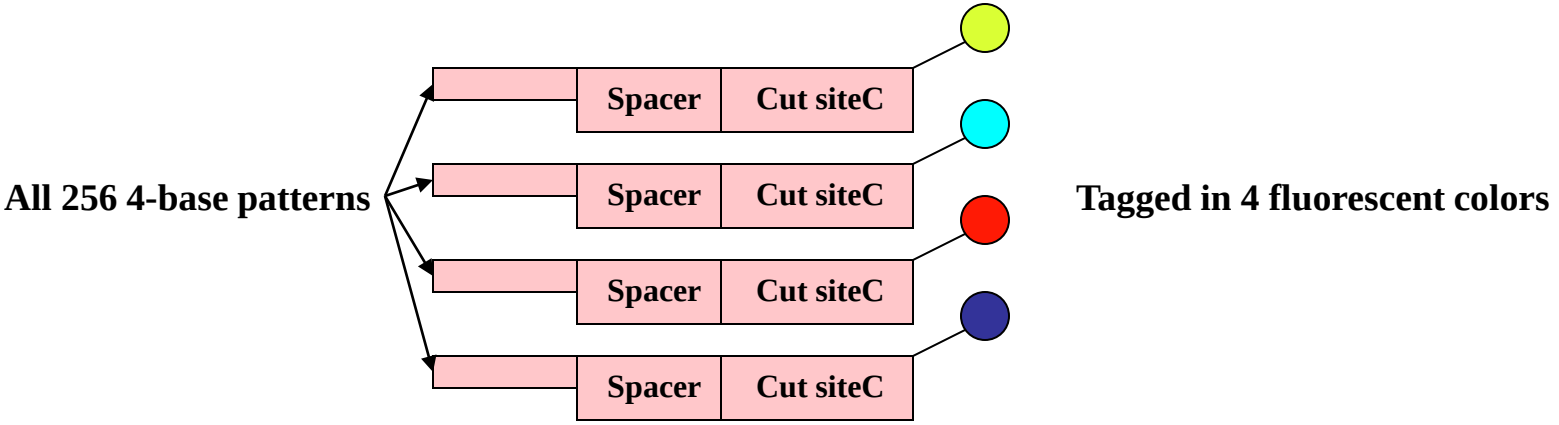
- **(Version 1: analysis) We will take a reading from a spot only when just one species of the population P hybridizes to it, i.e. only when the tag is represented just once in the original subcollection S of hybrids. The probability that a tag present in S is present only once is**

  $(1 - 10^{-6})^{10**6} = e^{-1}$, i.e. 37%. For much the same reason, the probability that a tag represented by a spot on the microarray does not appear in S is also 37%. So about 37% of the spots never hybridize, 37%hybridize to just one species of P, and 16% hybridize to 2 or more species drawn randomly from P.

MPSS: structure of the tag oligos and initiating adaptor

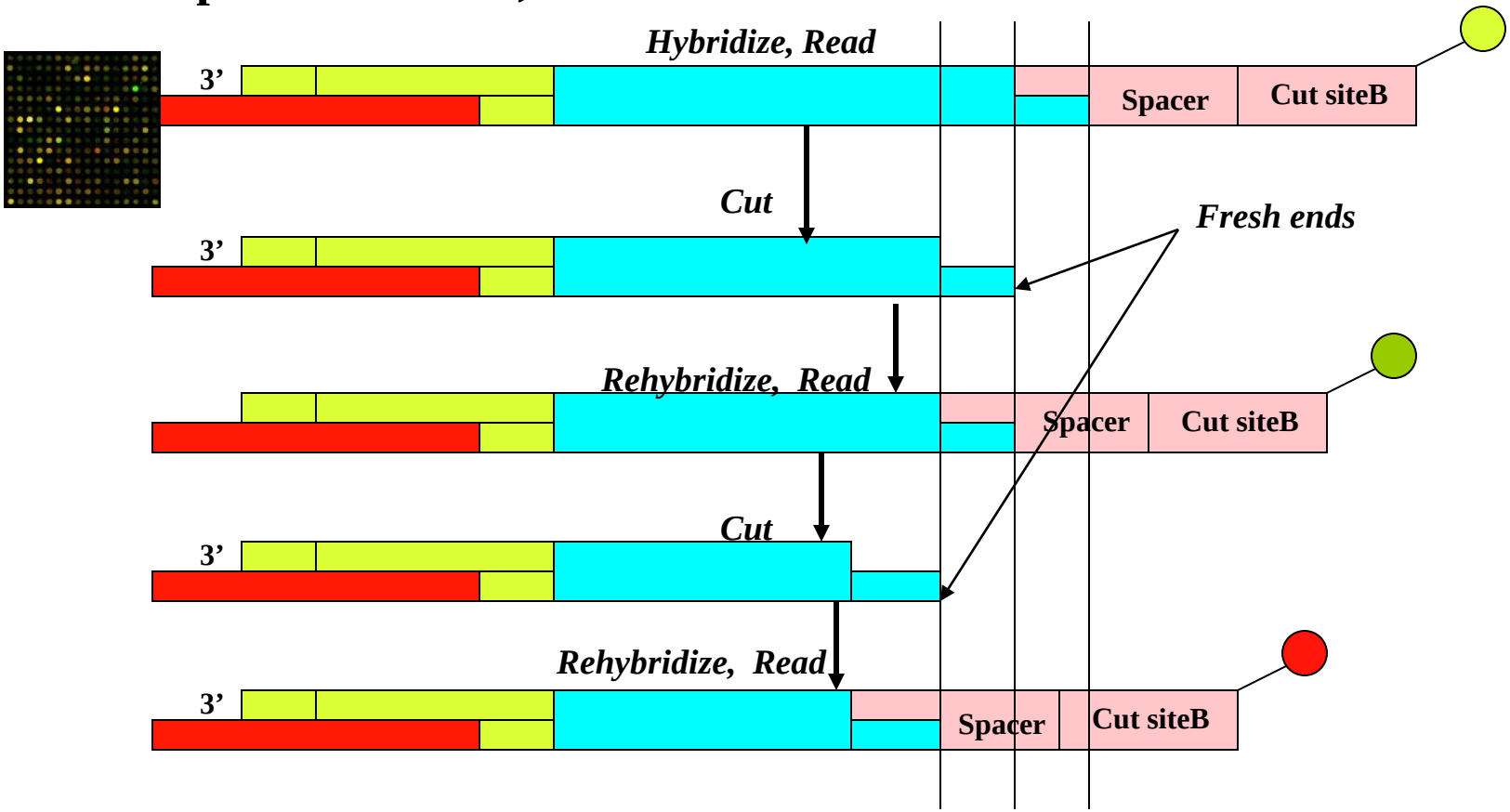


MPSS: structure of the (1,024) decoding adaptors



MPSS: sketch of the reading cycle

- **Repeatedly**
  - **Hybridize *decoding adaptors* to exposed end of the chip-bound oligos**
  - **Determine bases in positions near the end by observing fluorescence**
  - **Cut to expose fresh ends, read as often as desired.**



MPSS: improved reading efficiency (if genome known)

- **Each read cycle tells us (for each spot) which bases (or bases) is encountered at each position.**
- **If just one such is encountered at each position, we know that the tag of that position is attached to just one cDNA molecule in our sample.**
- **If just two (resp. 3) such are encountered at each position, we know that the tag of that position is attached to just two (resp. 3) cDNA molecules in our sample.**
- **If 4 such are encountered, the spot is unreadable.**
- **If there are n cDNAs in the collection being analyzed, and we ignore all cases in which more than one cDNA hybridizes to each spot, then $(\log_2 n)/2$ bases need to be decoded to characterize the hybridized base completely, since each decoded base contributes 2 bits of information. If we also allow cases in which 2 cDNAs hybridize to one spot, then $\log_2 n$ reads are needed, since each read contributes 1 bit. If 3 are allowed, then $2 \log_2 n$ are needed, since each read contributes just 1/2 bit.**
- **It is always best to allow up to 3 cDNAs at each spot, starting with an initial sample of 2.25 million tagged cDNAs, giving 1.37 million valid reads, with 11% of the spots empty and 19% unreadable.**

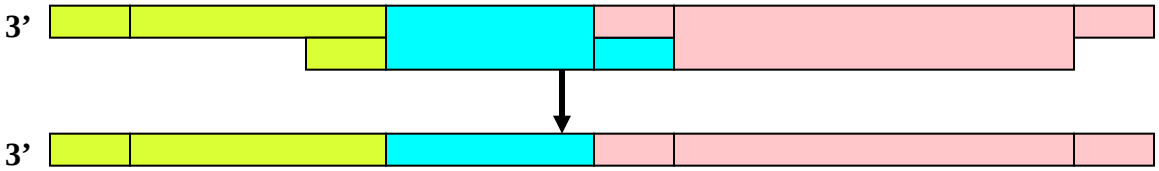
MPSS: result and some potential applications

- **A random sample of 1,370,000 cDNAs can be read in parallel using one chip and repeated enzymatic operations.**
- **Standard application: find cDNA expression levels.**
- **Potential application: census of bacteria and viruses.**

MPSS: getting statistics for the less common cDNAs

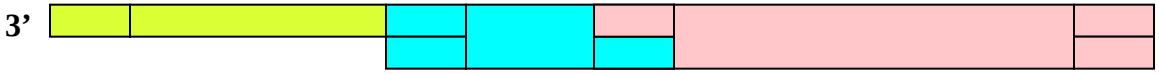
- **Once statistics have been gathered for the commonest cDNAs in the original sample of tagged cDNAs, all these cDNAs can be eliminated from the source pool of cDNAs by the method shown below. Then a new sample, contining only the next most common items, can be selected and processed. By repeating this process as often as necessary, the expression statistics for all the cDNAs can be understood.**

1. **Completely remove the undercut strand from all the tagged cDNAs.**



T7 Exonuclease

2. **Attach the known starts of all the common cDNAs, and polymerize, making the common cDNAs double-stranded.**



Common cDNAs only

3. **Erase all the double stranded DNA, leaving only the rare single-stranded cDNAs.**

Nuclease BAL-31

4. **Using their known ends, repolymerize the rare cDNAs, and then proceed as before.**

Rare cDNAs only