Basics

- **Molecules like rubber and plastic are arbitrarily long polymers of small molecules**

- **Think of them as *chains.***

- **A simple chain, with only one repeating element, carries no information other than its length.**
- **The key molecules of biology are *heteropolymers*, involving several different repeating elements.**
- **Think of them as necklaces of *different* beads.**
- **Such molecules are strings capable of**
  - **carrying indefinite amounts of memory,**
  - **provided that they can be copied faithfully.**

Three kinds of molecules play a role:

- **DNA: Beads on a wire - 4 kinds of beads -- the *memory***
- **RNA: Beads on a string - same 4 kinds of beads -- the *messages***
- **Proteins: Magnets on a string - 20 kinds : the *computing elements***
- **The 4 kinds of 'beads' in DNA and RNA, which might be called**
  - **-1,+1,-2,+2, are conventionally A,T,C,G**
- **+n and -n 'beads' attract each other weakly, stabilizing combinations like**
  - **-1-1+1-2+2+1+2... or more conventionally AATCGTG...**
  - **+1+1-1+2-2-1-2... TTAGCAC...**
- **So a molecule of Dna and RNA put into a ‘soup’ of A,T,C,G beads, with a suitable ‘zipper’ molecule will copy itself into its ‘dual’.**
  - **All the basic large-molecule chemistry is linear.**

Proteins from DNA

- **The sequence of steps that lead from DNA to proteins, and the molecular machines that perform them, are:**
  - **DNA --(transcriptase)-> RNA --(splicers)-> edited RNA --(ribosomes) -> proteins**
- **Various of these enzymes are commercially available**

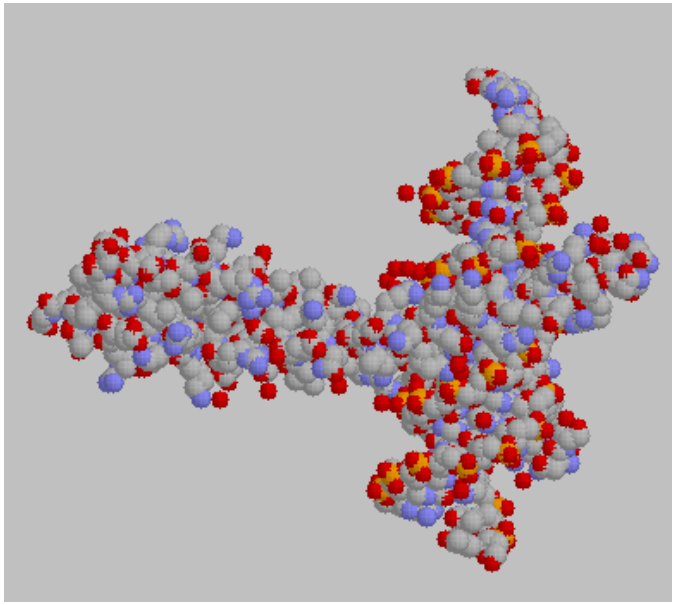
- **RNA editing and translation to proteins must still be done in cells ('vivo')**
- **DNA molecules can be very long (e.g. 2.5 Mb in *human*, 15Mb in *onion*)**
- **Unedited RNA molecules , copied from tagged sections of DNA, are typically 30-60Kb, reduced after editing to 1-2Kb, translated into proteins 300-600 beads long by 3-->1 *genetic code*. So only 3% of RNA survives editing, on average, and only half of DNA is ever transcribed to RNA.**

What living cells must do

- **A fourth basic activity of life is transfer of the *very large* DNA molecules into a new cell after they have been copied. (This is accomplished by a more complex sequence of steps.)**

- **A cell is a bubble containing DNA, RNA, and proteins, able to collect elements from the environment to make more, and able to transfer a copy of its DNA to a new cell copy once it has grown large enough. Cells can accumulate or secrete any kind of protein.**
  - **They must have polymerase to copy DNA, transcriptase to make RNA, ribosomes to make proteins, and an apparatus of division.**

- **Proteins exist within the cell as folded lumps**
  - **whose 'bumps' define their interactions**
  - **with each other and with DNA**
- **Viruses are simply protein-coated DNA**
  - **They cannot make proteins (no ribosomes)**
  - **And are therefore parasitic on cells for this**
  - **But they need no special division machinery**

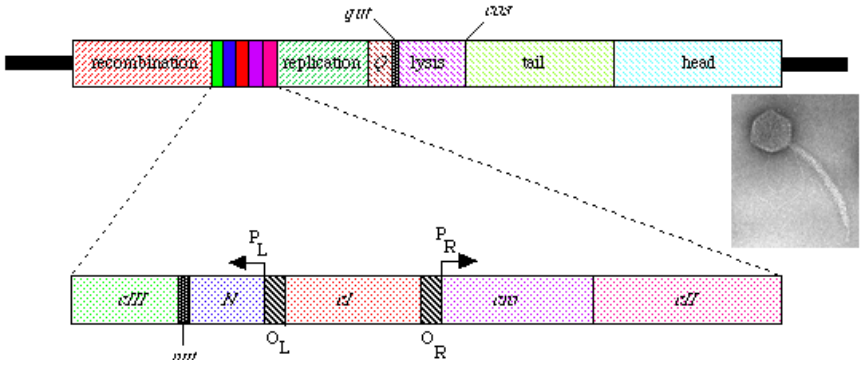


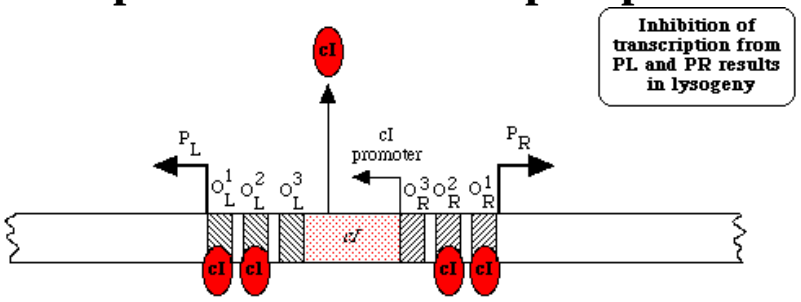
Means for reading DNA sequences are now available

- **The start of the smallpox genome:**
  - **1 atgattgtgt tattgatact atcgttagcg tgtacagcgt tcacctatcg cctgcaagga**
  - **61 tttaccaatg ccggtatagt agcgtataaa aatattcaag atgggaatga ggatgataat**
  - **121 attgtcttct cgccgtttgg ctattcgttt tctatgttta tgtcactatt gcctgcatca**
  - **181 ggtaatacta aagtagaatt attgaagact atggatttga gaaaaataga tctgggtcca**
  - **241 gcatttacag aattaatatc aggattagct aagccaaaaa catctaaata tacgtacact**
- **Start of a control protein gene determining male development in humans**
  - **241 agaaggcgaa ggctgcaggc gtgaggagct gtgactaatg agaattaaag gccatggatg**
  - **301 aagatgaatt tgaattgcag ccacaagagc caaactcatt ttttgatgga ataggagctg**
  - **361 atgctacaca catggatggt gatcagattg ttgtggaaat acaagaagca gtttttgttt**
  - **421 ctaatattgt ggattctgac ataactgtgc ataactttgt tcctgatgac ccagactcag**
  - **481 ttgtaatcca agatgttgtt gaagatgttg tcatagagga ggatgttcag tgctcagata**
- **These have the transparency and charm of hex code dumps**
  - **But can inspect for interspecies (e.g. human/mouse) and inter-individual differences (remote phylogeny)**

Control

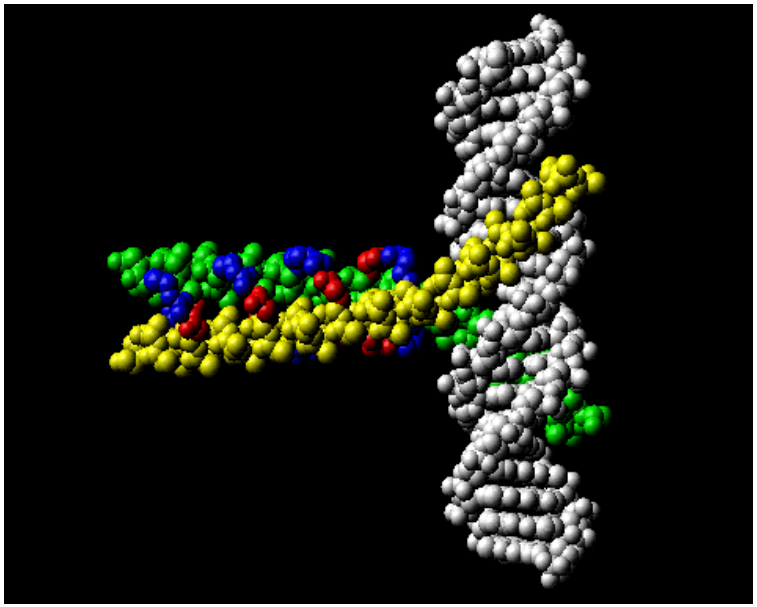
- **Cells detect changes in their external and internal environment as changing molecular concentrations, which may cause some proteins to bend or otherwise reconfigure slightly, changing some significant ‘bump’.**
- **The rate of production of each internal protein, and so its concentration, is controlled by the rate at which copying of its precursor RNAs from their DNA sources begin.**
- **For copying to begin, a ‘copying machine’ molecule (transcriptase) must be attracted to the start of the corresponding tagged section of DNA.**
- **The rate of copying can be changed greatly (e.g. *100) by an auxiliary protein which attaches itself to a particular section of DNA, either to attract or to block/repel the ‘copying machine’. Lambda-phage in E. Coli is a famous case, illustrating the biological implementation of a flip-flop.**





  - **48,502 bases, 71 genes.**

Control



- **Two ‘leucine zipper’ molecules will form a ‘clamp’ stably attracted to DNA**
- **Provided that the ‘bumps’ at the clamp ends hold to the DNA sequences at the points of attachment**
- **This sets up an attraction (or obstruction) locus for transcription start.**

• In thresholded logical terms:

output_protein present := control_1_present & control_2_present, or
output_protein present := not (control_1_present & control_2_present)

• This lets the set of genes present on the DNA function as a program whose execution state is defined by the presence/absence (level) of the corresponding proteins.
- ‘Programs’ have additive flavor of Markov-rule systems

A speculation: the cell as a molecular computer

- ***How might the cell function as a (universal) molecular computer*?**

- **Use (e.g.) 12 proteins, present or absent, to define a master 'state': 4096 states.**
- **Sense the external and internal environment by activating/inactivating, then producing/eliminating signal proteins. (As many as appropriate)**

- **The state proteins catalyze their own continuance (as in lambda), except in presence of a periodically varying 'clock' protein, which acts with an auxiliary pair of control proteins to realize**
  - **state_n_present := if clock_present &aux_n1_ present then true**
  - **elseif clock_present &aux_n1_ absent then false else state_n_present end if**

- **The aux_nj _present/aux_n1_ absent protein values, and any auxiliary output_present values,can be defined by boolean expressions in the state_n_present and signal_j_present values. These boolean expressions can be realized using the '&' and 'not-&' operations described previously.**

The cell as a molecular computer, II

- **The machinery needed consists of:**
  - **Ribosomes for protein synthesis (53 proteins, 3 RNAs)**
  - **Auxiliary polymerases for DNA and RNA, problem sensing, DNA repair**
  - **Houskeeping: energy, synthesis of bases (‘beads’), amimo acids (protein ‘beads’): estimate 850 + 130**
  - **Detector molecules for external and internal conditions**
  - **Structural proteins for control and execution of the division cycle**
  - **Various small molecules, cell wall maintenance and coating**
  - **Control proteins (e.g. E. coli has about 700 known)**
- **E. Coli genome is 4,639,221 bp, 4.300 known genes; human has 2.5 Gb. 37,000 known/suspected genes. E. Coli detects at least 30 conditions, plus overheating.Yeast about 12Mb., about 6,000 genes.**
- **Our conjectured 4K-state cellular computer is probably complex enough to sustain multicellular life and cell specialization.(must react to presence of other cells)**
- **Cellular systems are inherently parallel at all levels from their Markov-condition programming on up.**

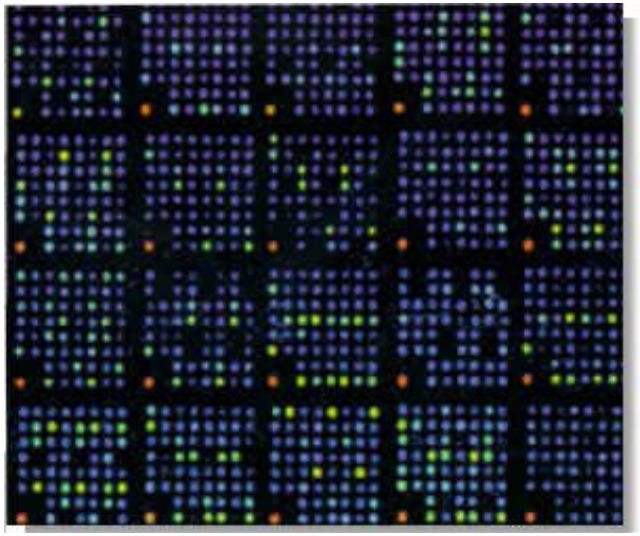
Programmable systems can have bugs

- **A bug in the program of a germ cell leads to failure of development or a genetic disease, of which there should be at least 37,000**
- **About 2 million mutations have accumulated (in humans)in the last 100,000 years: 4000 generations, so 500 per generation, 10 per division.**
- **Wandering retroviruses install themselves in the genome, and spread mutations. E.g. Rous sarcoma virus contains a mammalian growth gene.**
- **Bugs in the program of a single cell are not important, unless they cause uncontrolled growth (cancer) by**
  - **(1) injury to something in the DNA copying/repair/segregation pathway, raising the mutation rate, and producing progressive collapse;**
  - **(2) injury to some growth control factor, producing uncontrolled growth;**
  - **(3) injury to an adhesion-molecule gene (easier metastasis, producing injury to some important tissue);**
  - **(4) injury to some condition sensitivity gene, allowing cells to survive in environments in which they would normally sense trouble and die.**

The genomes of tumor cells are damaged, often quite badly

- **Numerous genes, even whole chromosome sections. may be deleted or duplicated**
- **Close inspection of tumor-cell genomes may provide useful hints for choice of therapy and prognosis.**
  - **New tools are becoming available for such inspection**
  - **E.g. DNA microarrays**
  - **Can inspect DNA or RNA**
    - **Via reverse transcription to DNA**
  - **Photolithographic synthesis**
    - **up to 100,000 spots, oligos 70 long**
    - **Probes synthesized directly onto glass**
    - **treatment with each of (A,C,G,T) serially**
    - **Apply series of lithographic masks.**
  - **Array is hybridized to mix of DNA fragments**
    - **obtained from DNA to be analyzed (tagged red)**
    - **And comparison DNA (tagged green)**
  - **Piece complementary to each oligo will bind**
    - **at the corresponding array spot**

AFFYMETRIX®



NimbleGen™ SYSTEMS, INC.

The chemistry, magic bug juices, and procedures used

- **Chemistry = synthesized, e.g. oligonucleotides**

Synthesis Scale: 1micromol

Modification	Base Charge	Setup Charge	Modification Charge**	desalt	OPC	HPLC
None	$2.50	$0.00	$0.00	$5.00	$5.00	$80.00
Biotin	$2.50	$0.00	$150.00	$5.00	$5.00	$80.00
Inosine	$2.50	$0.00	$45.00	$5.00	$5.00	$80.00
Uridine	$2.50	$0.00	$45.00	$5.00	$5.00	$80.00
5'Phosphorylation	$2.50	$0.00	$45.00	$5.00	$5.00	$80.00
Phosphorothioate	$2.50	$0.00	$0.00	$5.00	$5.00	$80.00

So a micromole of TTAAGGCGA will cost you $22.50

- **Bug juices = cell extracts, e.g. reverse transcriptase, DNA Ligase**
  - **Proteins still can't be synthesized, but their DNA can be inserted into convenient hosts and harvested (bacteria, yeast, mouse, etc.)**

Order This | Search Catalog | View Ice

Online Catalog

Restriction Endonucleases

NEW ENGLAND BioLabs Inc.

Aat II

#R0117S	500 units	$50 (USA)
#R0117L	2,500 units	$200 (USA)

5´... G A C G T^C ... 3´
3´... C^T G C A G ... 5´

Aat II	Bmt I	BstY I	Kas I	Pvu I
Acc65 I	Bpm I	BstZ17 I	Kpn I	Pvu II
Acc I	Bpu10 I	Bsu36 I	Mbo I	Rsa I
Aci I	BpuE I	Btg I	Mbo II	Rsr II
Acl I	BsaA I	Bts I	Mfe I	Sac I
Acu I	BsaB I	Cac8 I	Mlu I	Sac II
Afe I	BsaH I	Cla I	Mly I	Sal I
Afl II	Bsa I	CviA II	Mme I	Sap I
Afl III	BsaJ I	Dde I	Mnl I	Sau3A I
Age I	BsaW I	Dpn I	Msc I	Sau96 I
Ahd I	BsaX I	Dpn II	Mse I	Sbf I

- **The polymerase chain reaction allows any section of DNA addressed by short sequences at its two ends to be amplified exponentially**
  - **Much the same technology allows fast DNA sequencing**

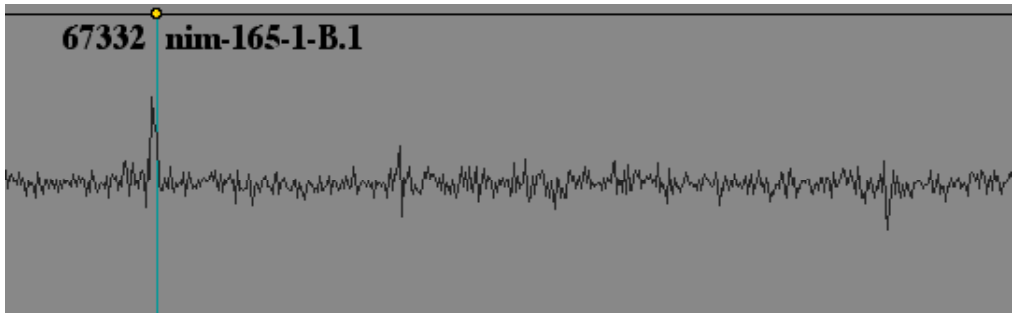
A DNA Lesion Assay (Wigler, Lucito, et al, CSHL)

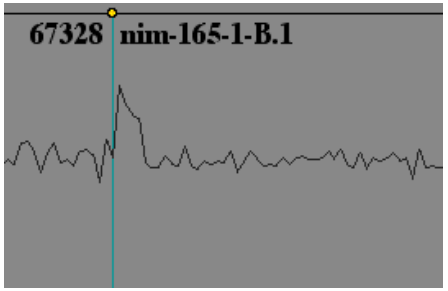
- **Exploit fact that normal human genome sequence is now known**

- **Approx. 3% DNA subsampling by PCR to reduce random hybridization noise**
  - **Cut with '6-cutter' restriction enzyme averaging 1 cut per 4KB: about 500K cuts**
  - **Position of cuts known; known 5-base sequences at ends at cut allow PCR**
  - **Selective PCR amplification advantage of shorter pieces tends to concentrate product among those pieces of length < 1500Bp**
  - **These pieces are known in advance from genome sequence;**
    - **there are about 135,000 of these; select 85,000 with highly distinctive 70-bp long internal sequences**
- **Tag sequences with red or green phosphorescent marker at an end**
- **Prepare NimbleGene chip with 85,000 complementary sequences**
- **Comparative hybridization detects changes in DNA**
  - **normal/normal or normal/tumor comparisons**

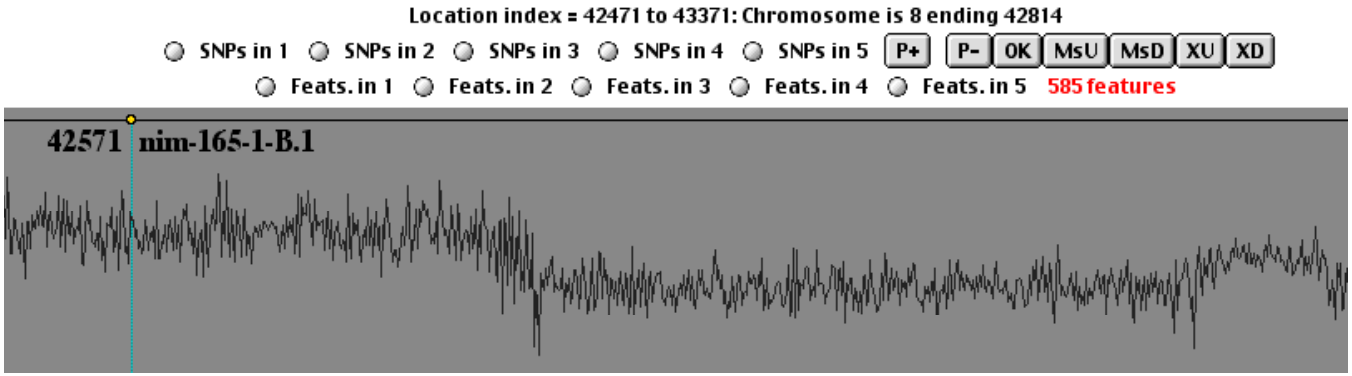
Features detected

- **Aberrant behavior of single or adjacent probes**





  - **could be point mutation adding an internal cut site or deleting one at end**
  - **could be copy-number change affecting region spanning part of 70-base probe region (or pair of regions)**
- **Large-scale change of signal read from extensive genome section**
  - **Could be deletion of amplification of chromosome section**
    - **Possibly within developing clone**

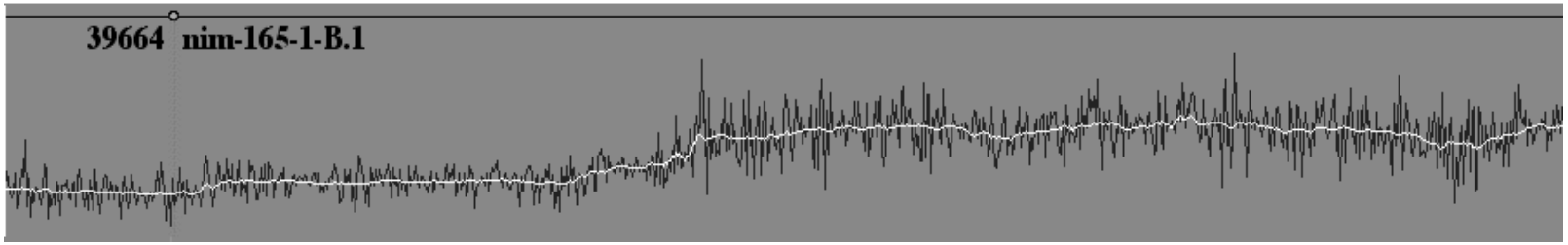


Signal detection issues

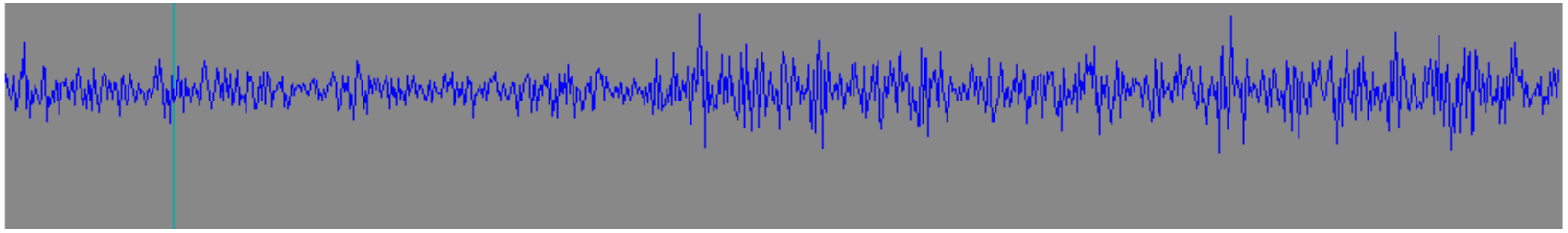
- **Given a signal corrupted by noise:**
  - **use the assumed characteristics of the signal and the noise to separate these as cleanly as possible.**
  - **signal reconstruction is optimal when difference of reconstructed signal and observed data can be seen as pure noise carrying no signal.**
- **In the present case:**
  - ***signal* is arises from copy numbers jumping between discrete values at a limited number of places along the genome (hence basically a step function)**
  - **plus a number of 'small lesions', affecting just a few probes , perhaps just 1**
  - ***noise* is assumed to be additive Gaussian**
- **Reconstruction heuristic:**
  - **calculate average and standard deviation of the data over two 16-probe-wide sliding windows to the left and right of each point x;**
  - **Select l_value or r-value from smaller of**
    - **(lr_average - measured_signal_at_x)/ lr_standard_deviation**

Results

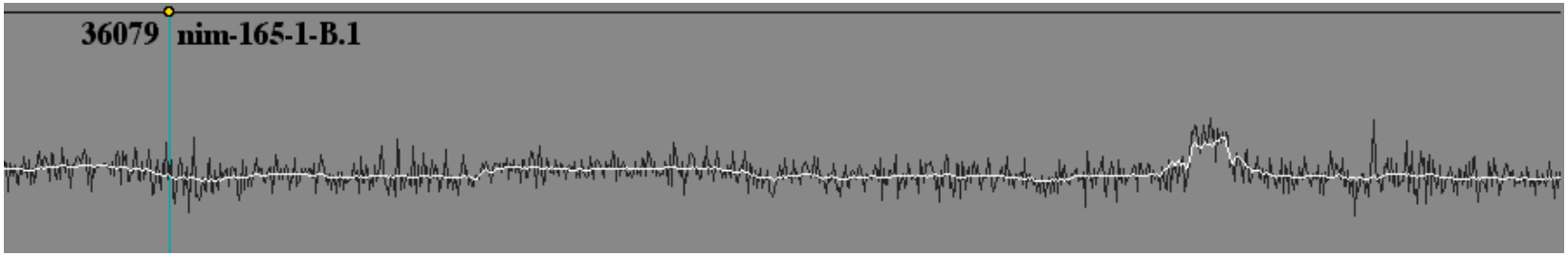
- **Mollified signal versus original**



- **‘Residual noise’**

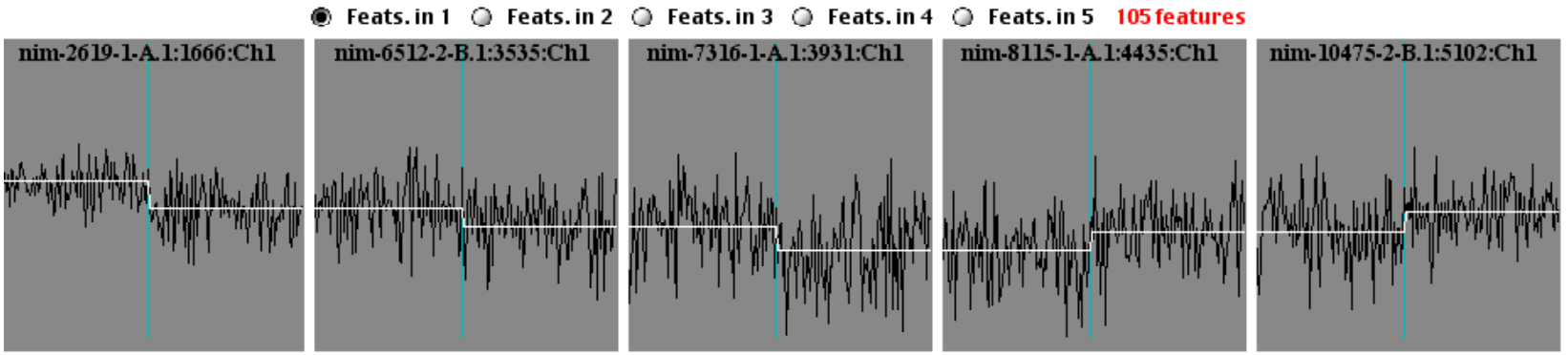


- **Quality of feature tracking**

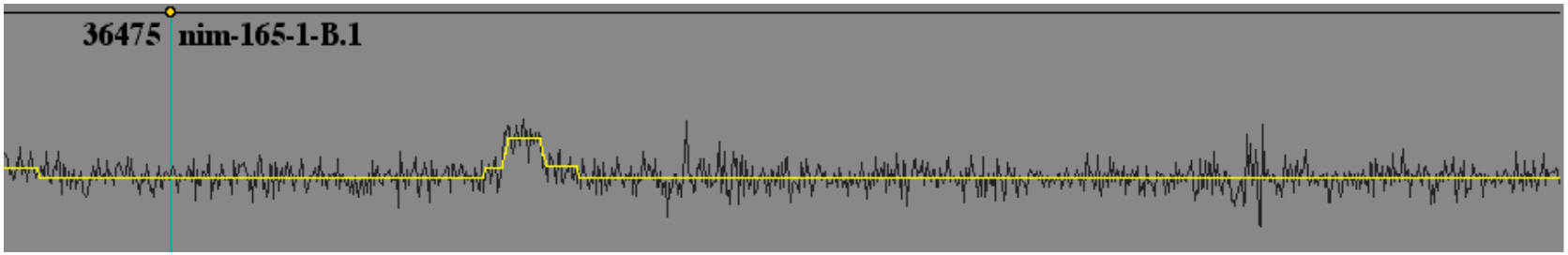


More drastic flattening needed for rapid scan

- **Several hundred features detected per run**
  - **will be manually examined for possible follow-up**



- **Second heuristic: track forward thru mollified data, finding zones where max - min remains below threshold; take average in these zones**

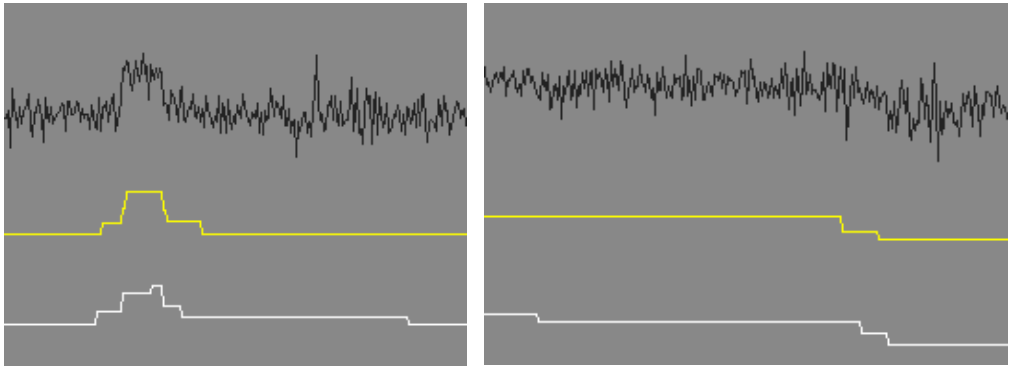


- **‘narrow’ features remain in ‘residual noise’ and can be detected there**

Performance assessment

- **When there is an underlying ground truth, given by a numerical function *gt(i)* of a real or integer parameter i:**
  - **In case at hand *gt(i)* is the 'true' copy number of the bases in the range covered by probe i**
  - **Measurement model is: *meas(i) = F(gt(i),noise(i))***
    - **for lack of knowledge we take F = sum and noise = gaussian**
  - **Reconstruction algorithm should minimize cost value C(*gt,reconstructed_truth*)**
  - **In the case at hand C = cost is number of missed features**
  - **Stability test: assume one interpretation is ground truth; assess sensitivity to noise, regenerated artificially to match noise statistics of measurement**

- **Results: 5 sample test runs:**



  - **small feature error rates:**
    - **0.27-; 0.46+**
    - **0.22-; 0.25+**
    - **0.29-; 0.19+**
    - **0.26-; 0.42+**
    - **0.33-; 0.19+**

wide feature error rates:

0.09-; 0.29+
0.19-; 0.08+
0.16-; 0.14+
0.19-; 0.11+
0.24-; 0.05+;