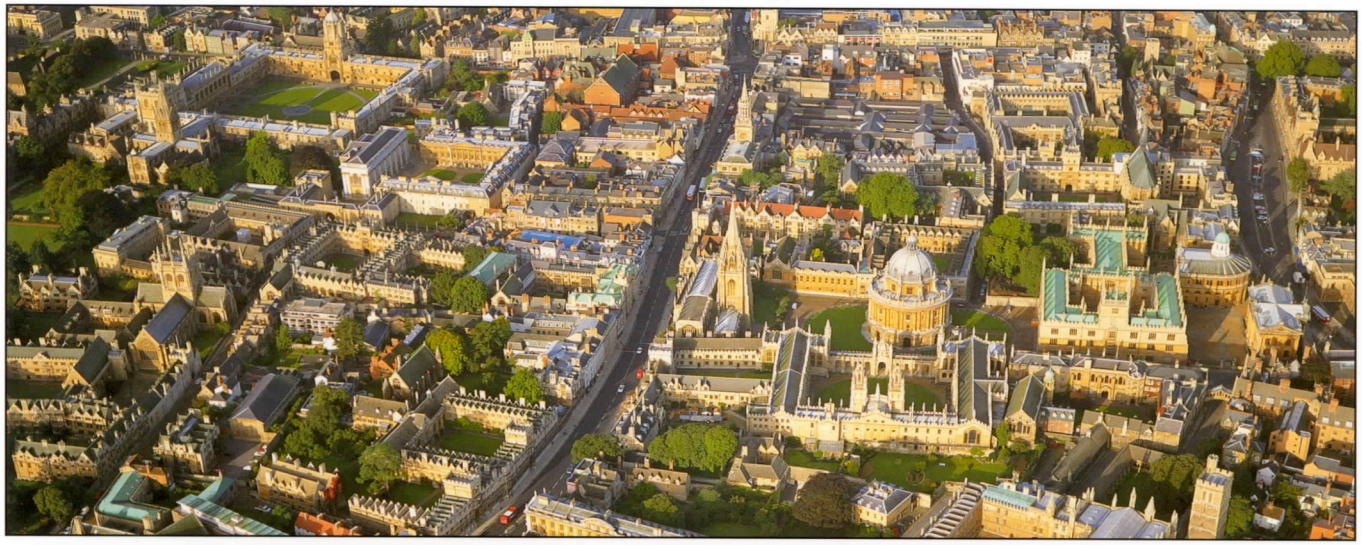
New sequencing technologies

Salisbury District Hospital, UK

18th July 2006



Laboratory of Hagan Bayley
University of Oxford

US National Institutes of Health

$1000 human genome by 2015

$100,000 as intermediate goal by 2010

Why a $1000 genome?

personal genome
- disease prevention
- diagnosis
- treatment

comparative human genomics

identification of pathogens

fundamental science
- cancer progression- sequence complete genomes from cancer cells

single cell sequencing
- microorganisms dead/ unculturable
- cancer cells
- individual chromosomes

Present costs ...

During the Human Genome Project the cost of sequencing dropped from $10/ base (1985) to $0.1/ base at completed draft (2001)

Finished sequences of human chromosomes still being published, e.g. chromosome 17 (Nature- April 2006). Cost now approaching $0.001/raw base (2005) on 384-capillary machine, which produces 24 nt/s

No change in the basic principles of sequencing during this period

- automation
- miniaturization- less reagents
- new biochemistry and software for sequence assembly
- local and collaborative management

$1000 human genome is $0.0000003/ base
cf. the $5000 per person per year spent on health care in the USA
$100,000 genome is $0.00003/ base, i.e. at least 30X cheaper than present costs

If a human genome were $1000: a bacterial genome would be $1! - doesn't quite scale, but an interesting thought

Progress in new sequencing technologies

1. Improved electrophoretic sequencing

2. Sequencing by hybridization (to chips)

3. Cyclic array sequencing

4. Single cell/ chromosome sequencing

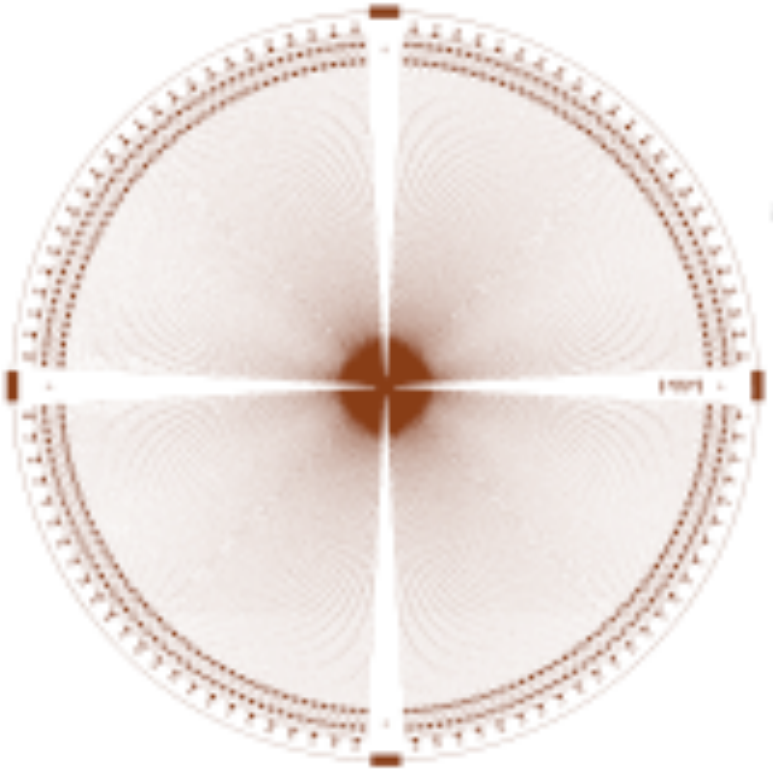
5. Single molecule techniques

J. Shendure, R. D. Mitra, C. Varma and G. M. Church (2004). "Advanced sequencing technologies: methods and goals." Nature Rev.Genet. **5**: 335-344.

Electrophoretic sequencing

Practical long sequence reads
1970s

- Maxam-Gilbert chemical cleavage
- Sanger sequencing- chain termination
- Electrophoretic separation

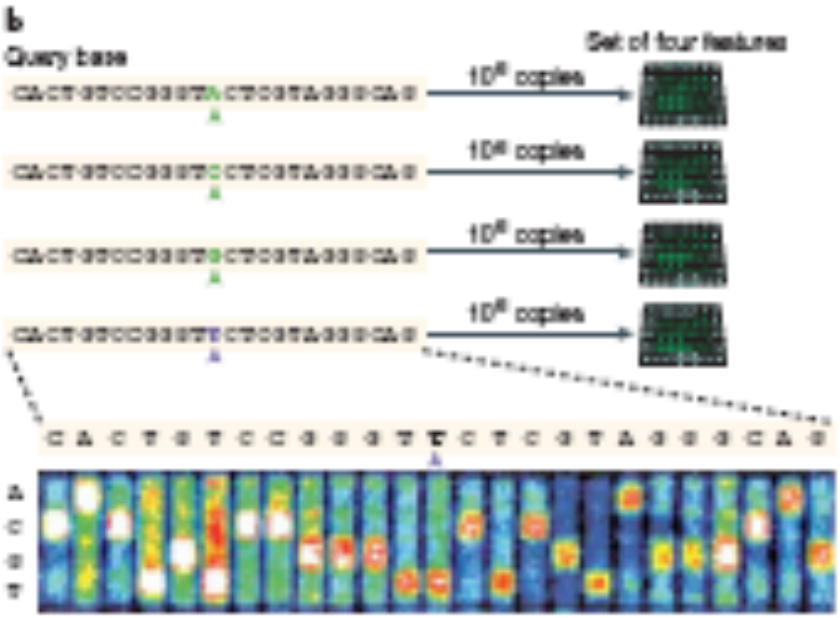


Amplify, purify, sequence on a single device

... but improved electrophoretic sequencing unlikely to give 10,000- 100,000-fold improvement for $1000 genome- but $100,000 genome a possibility...

Sequencing by hybridization

hybridization target on chip.... or ....
oligos on chip-- (e.g. 25-mers)
    Affymatrix (HIV chip)
    Perlegen (chromosome 21 chip)
requires amplification of region to be sequenced
    problems with similarities and repeats
rapid re-sequencing of specific genomic subsequences may be a strength of this approach



Cyclic array

requires amplification of separate genomic fragments-

thousands to billions of samples in parallel

454 Life Sciences: picoliter wells

polony- single molecule amplification with restricted diffusion (emulsion or gel)

each cycle only examines one or a few bases but massively parallel

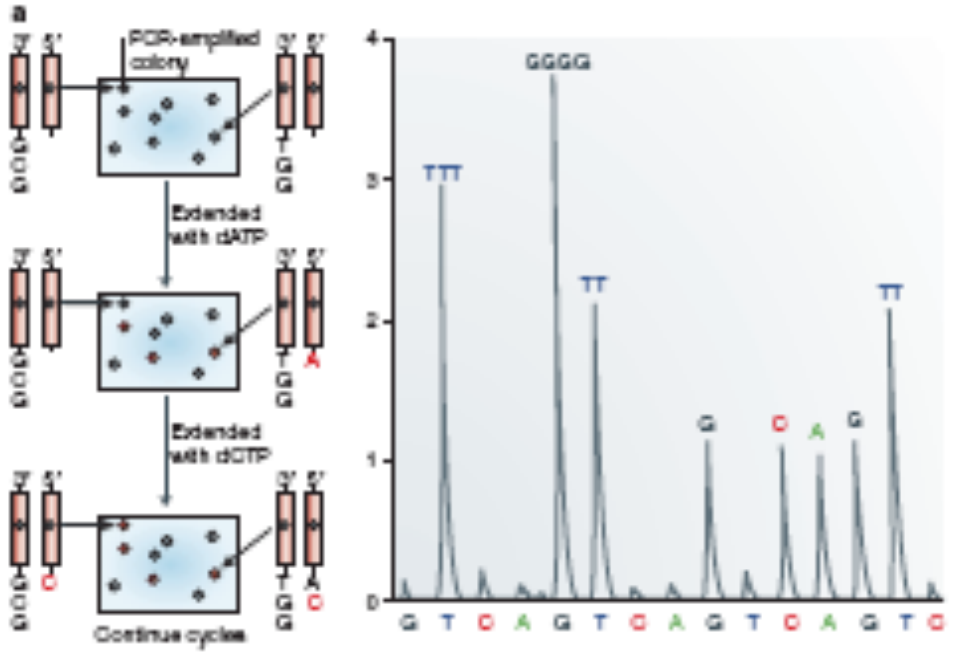
current approaches include (aka sequencing by synthesis)

pyrosequencing

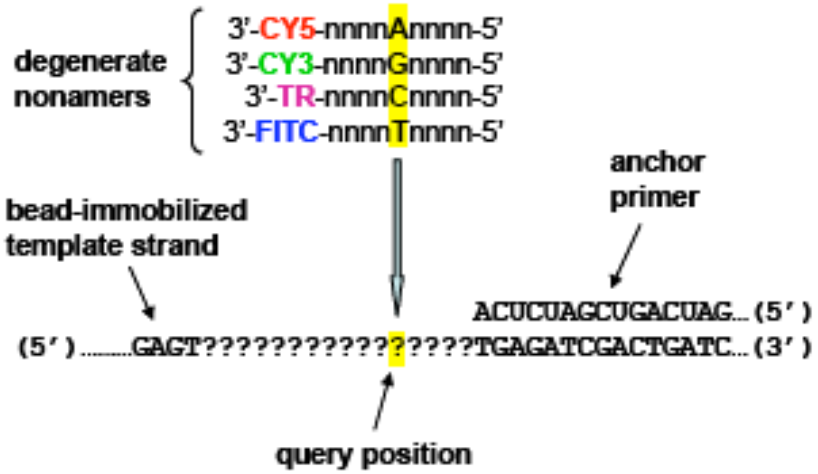
fluorescence in situ sequencing (so far <10 bases)

homopolymer problems- here reversible terminators might help

Could achieve $100,000 human genome



Shendure et al. 2005



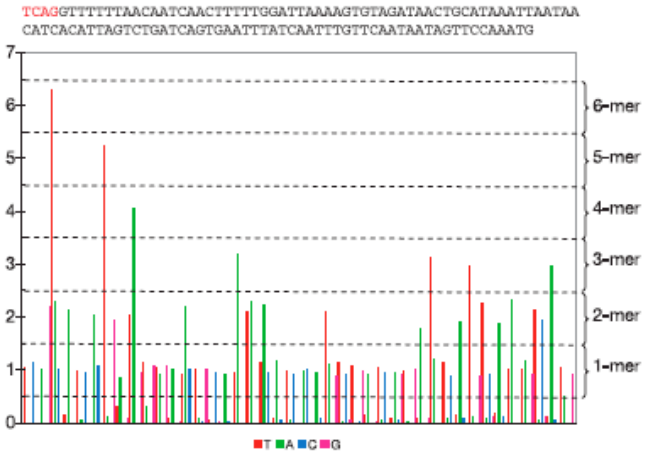
Inexpensive epifluorescence microscope

Cell-free library construction: single DNA molecules were amplified in parallel and captured by 1-micrometer beads by emulsion PCR

Millions of beads were immobilized in a polyacrylamide gel and subjected to automated cycles of sequencing by ligation and four-color imaging- only 26 bp per polony but many polonies

Fully automated. 60 h instrument run. Two runs required for entire E. coli genome. 71% high quality coverage (4X). Cost per base was roughly one-ninth as much as that of conventional sequencing. estimate $ 0.0001/ base

Shendure, J., G. J. Porreca, et al. (2005). "Accurate multiplex polony sequencing of an evolved bacterial genome." Science **309**(5741): 1728-32.



Margulies et al. 2005

Emulsion/ bead method for DNA amplification (10 h from genomic DNA)

Place in novel fiber-optic slide with individual wells

Pyrosequencing protocol optimized for solid support- reads of more than 100 bases

Advanced software for base calling

25 million bases, at 99% or better accuracy, in one 4h run.

Mycoplasma genitalium genome (small 0.5 Mb) with 96% coverage at 99.96% accuracy in one run 4h of the machine.

Margulies, M., M. Egholm, et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-80.

Single cell/ chromosome sequencing

whole-genome amplification

why useful?

- dead cells
- unculturable cells
- haplotypes/ diploid genotype
- splice forms- single RNA amplification in this case

Zhang et al. 2006

Real-time isothermal amplification to form polymerase clones (plones) from the DNA of single cells.

Whole-genome shotgun sequencing of Prochlorococcus MIT9312 plones showed 62% coverage of the genome from one plone at a sequencing depth of 3.5x, and 66% coverage from a second plone at a depth of 4.7 x.

Genomic regions not revealed in the initial round of sequencing are recovered by sequencing PCR amplicons derived from plonal DNA.

The mutation rate in single-cell amplification is $<2 \times 10^{5}$, better than that of current genome sequencing standards.

Note: Rapid sequencing is required to get the most out of this approach

Zhang, K., A. C. Martiny, et al. (2006). "Sequencing genomes from single cells by polymerase cloning." Nat Biotechnol **24**(6): 680-6.

Single molecule sequencing- advantages

No amplification required

cloning, PCR- costly and prone to errors, bias etc

... or less starting material needed if amplification is required

e.g. from a single cell

Haplotyping

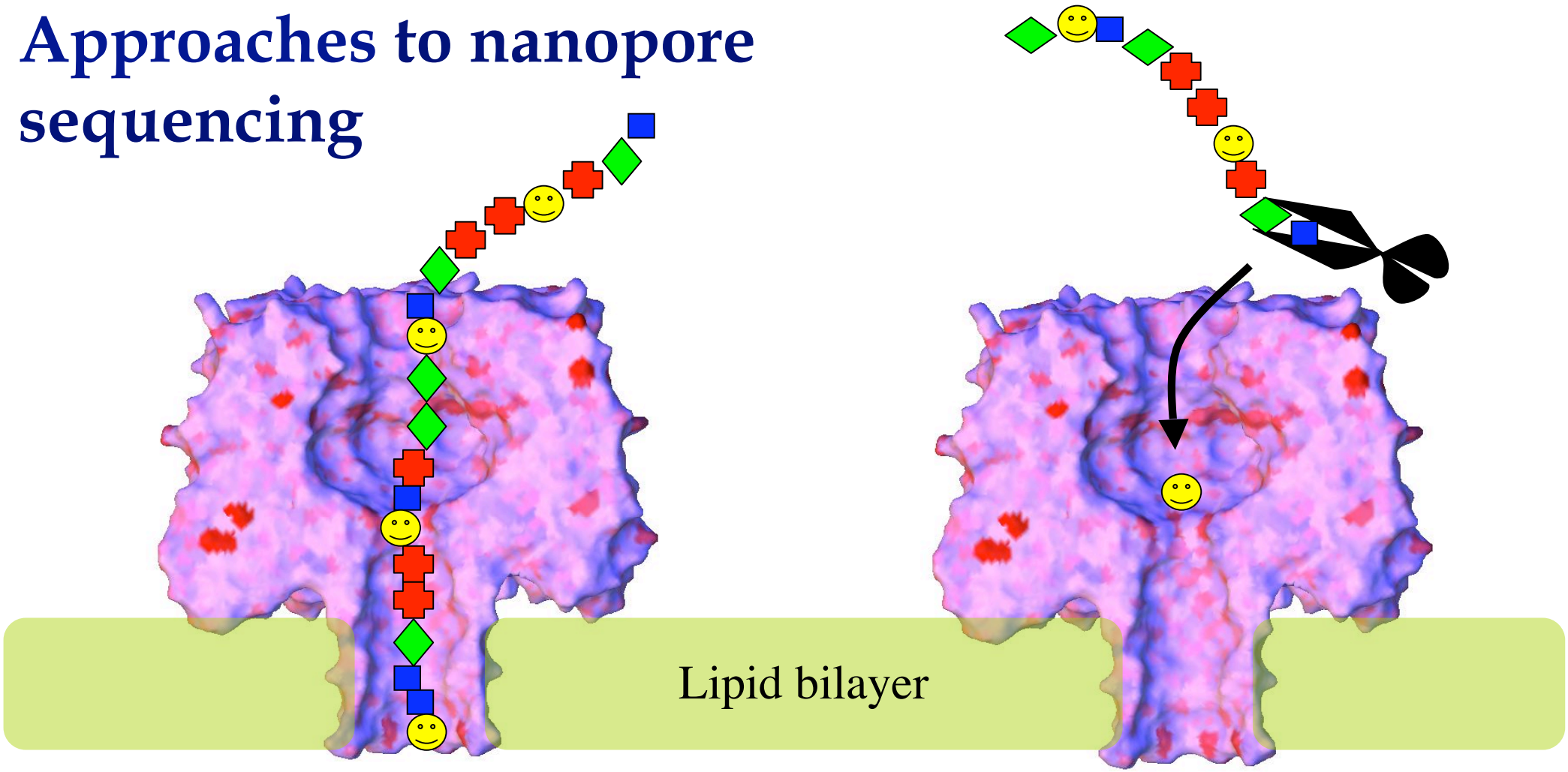
RNA- splice forms

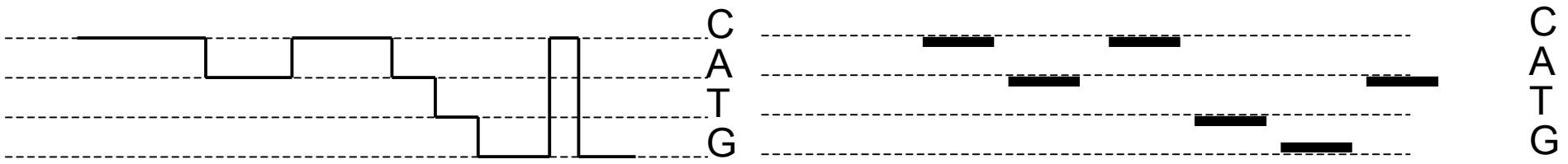
Approaches

- one-at-a-time base extensions with fluorescent probes of various kinds
  - several groups/ companies
  - cannot go out of phase with a single molecule
- polymerases in real time are a possibility too

- nanopore
  - very rapid
  - little or no sample preparation
    - need to amplify (but to a lesser extent) if a single cell
  - reagent-free

Approaches to nanopore sequencing





… unmodified DNA sequencing …

Why nanopore DNA sequencing?

Size of human genome: 6,000,000,000 bases (1000 Russian novels)

At 1 millisecond per base (likely ultimate speed):

6,000,000 sec = 100,000 min = 1700 hours = 70 days

1000 pores in parallel, 10 times coverage: 0.7 days

Bacterial genome: 20 sec

Current operation is 24 nt/sec per device

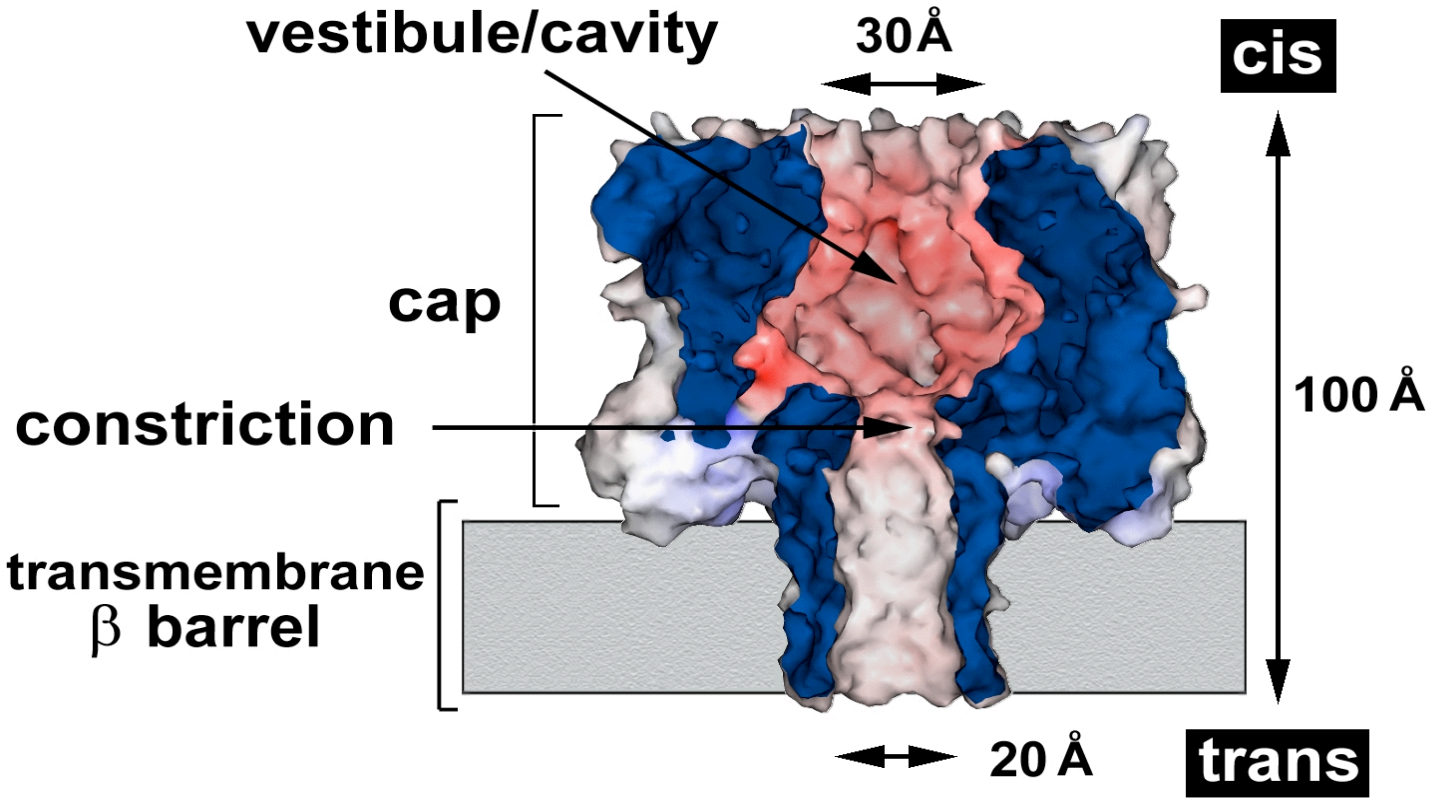
For $1000 genome need to increase throughput at least 20,000-fold at same unit cost

Church calculates that a resequencing device would need to do 60 bp reads with 99.7% accuracy at 500,000 nt/sec

A 1000-channel nanopore device with no significant pauses between strands could do this

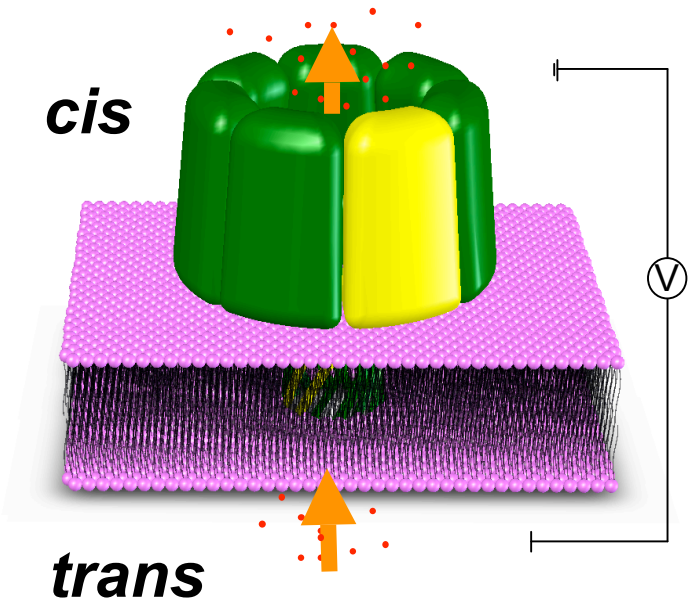
Chemistry at the single-molecule level:
using a transmembrane protein pore as a nanoreactor

α-hemolysin pore as a nanoreactor

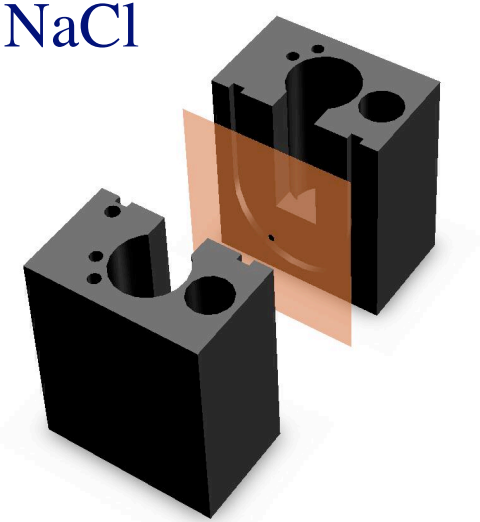


Current measurements are made by single channel electrical recording in planar bilayers

Perturbations of the current tell us what is happening inside the pore, we see ***individual*** interactions or reaction steps for ***single*** molecules



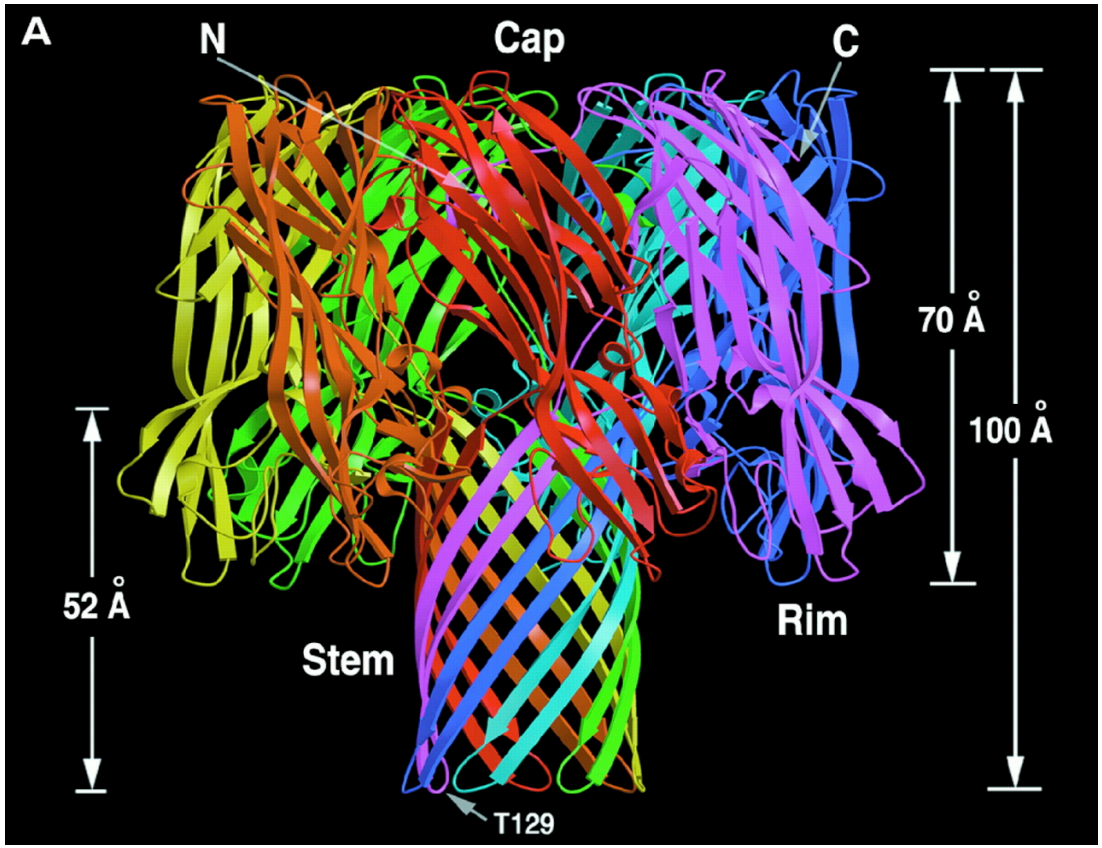
ΔV = 40 mV
700 pS = 30 pA
1 M NaCl

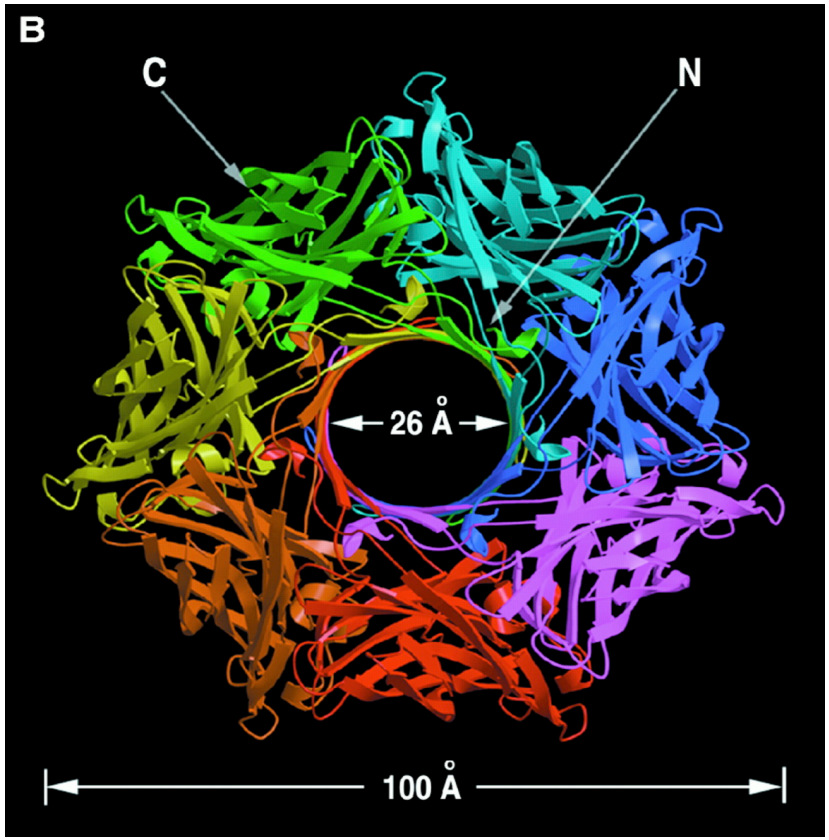


In this talk:
various ΔV, salt, filtering, timescales etc.

α-Hemolysin (α-toxin)

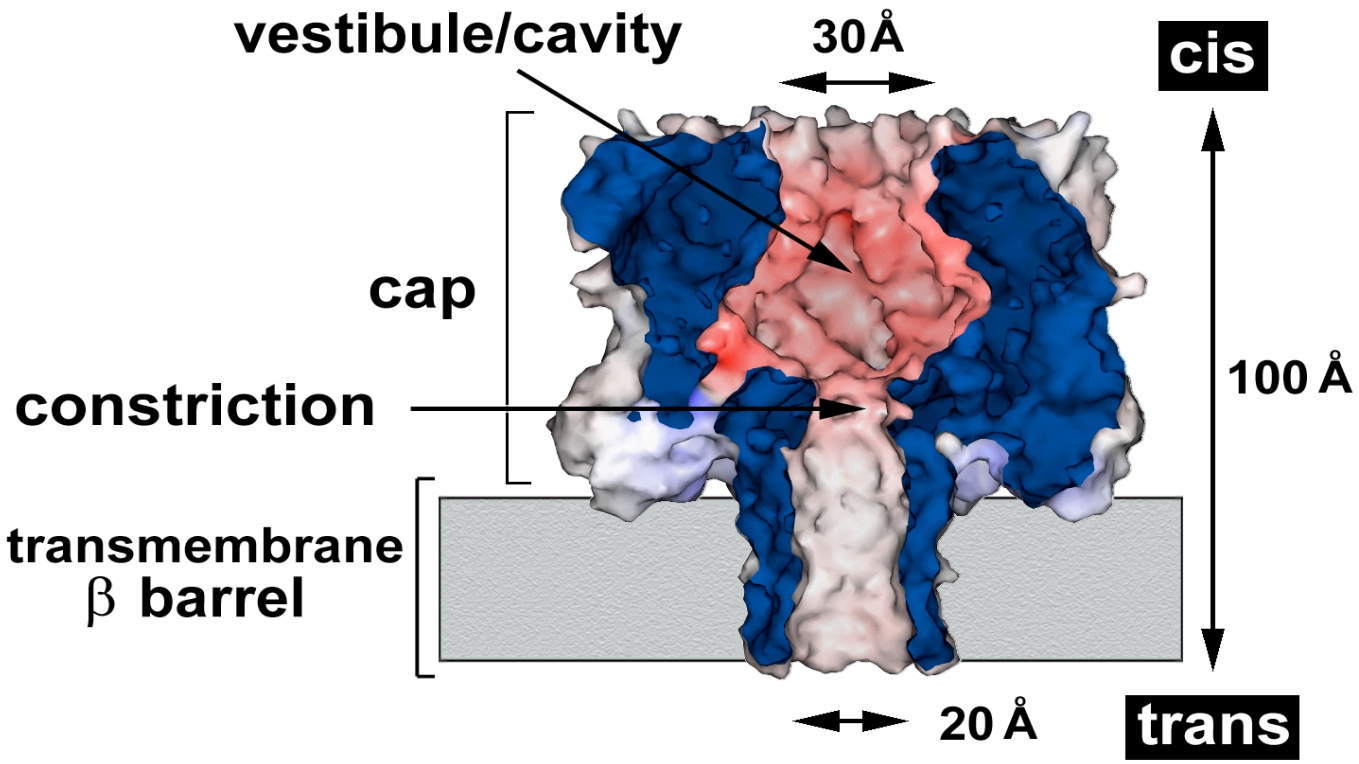
- secreted by Staphylococcus aureus
- 293 amino acid polypeptide
- water-soluble monomer
- heptameric transmembrane pore



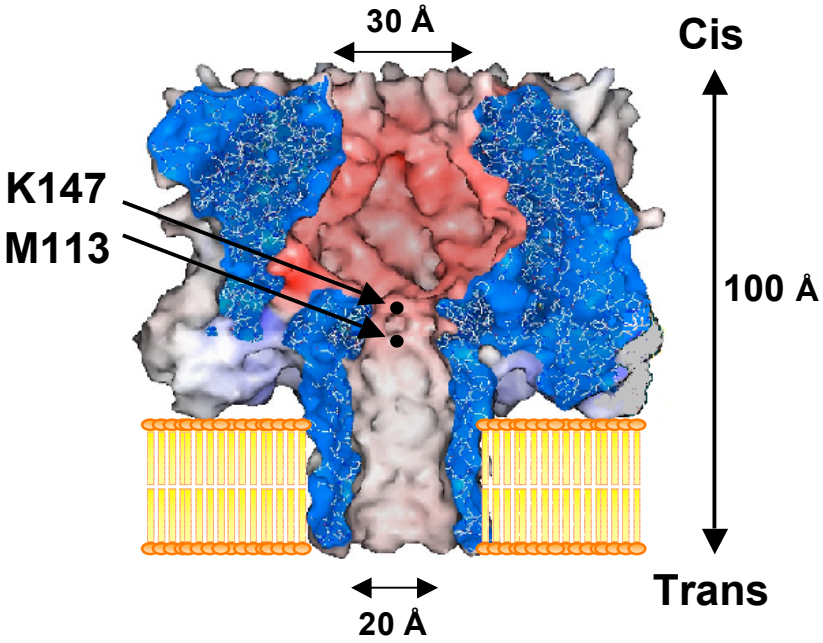


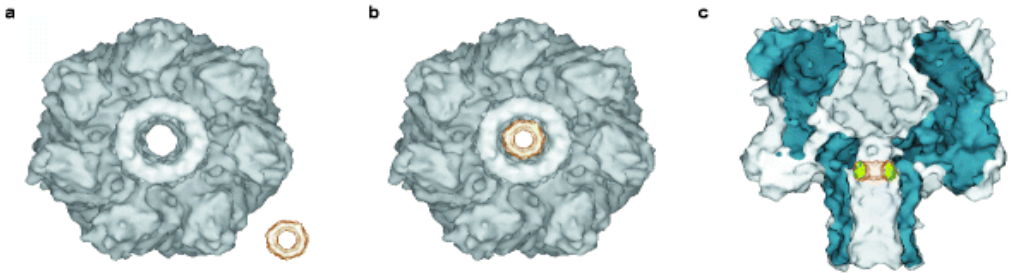
Eric Gouaux and colleagues solved the structure of the α-hemolysin pore at the University of Chicago

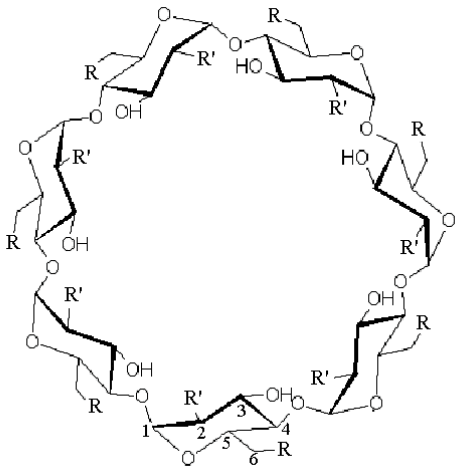
Engineered versions of the α-hemolysin pore can be used for the detection of analytes with exquisite discrimination

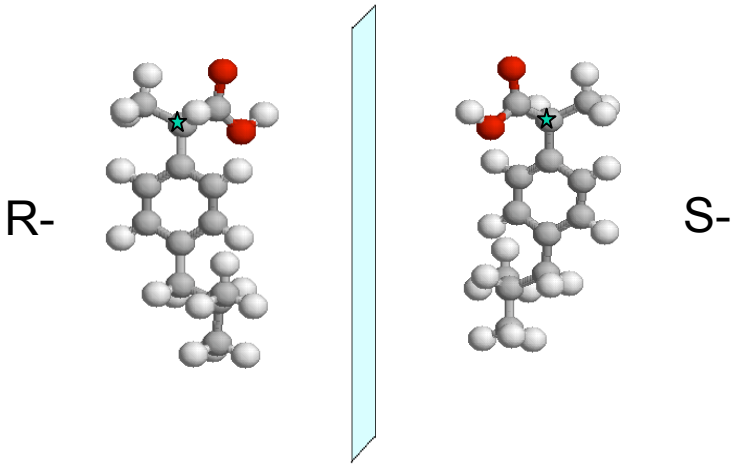


Using non-covalent molecular adapters for detecting organic molecules

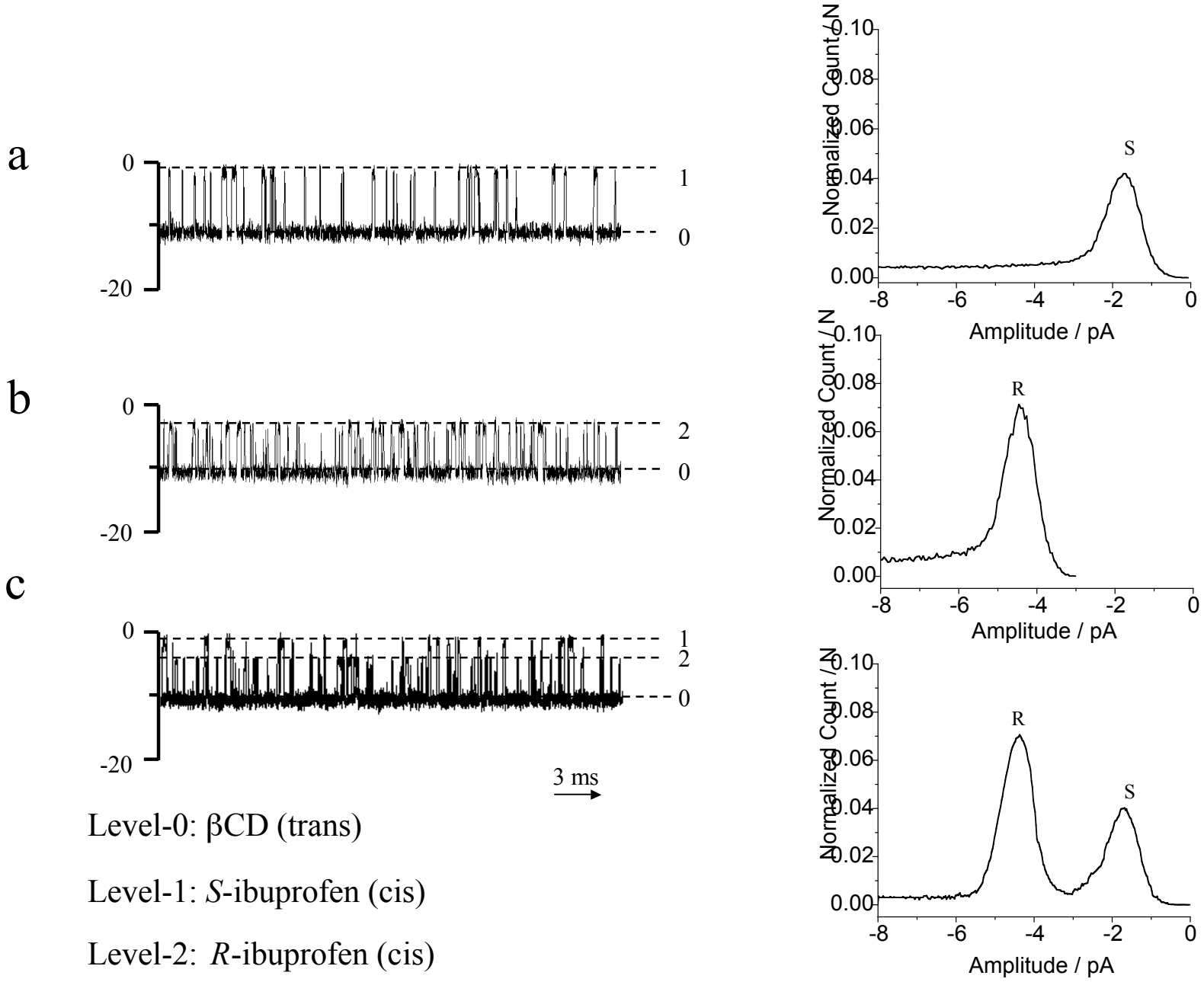




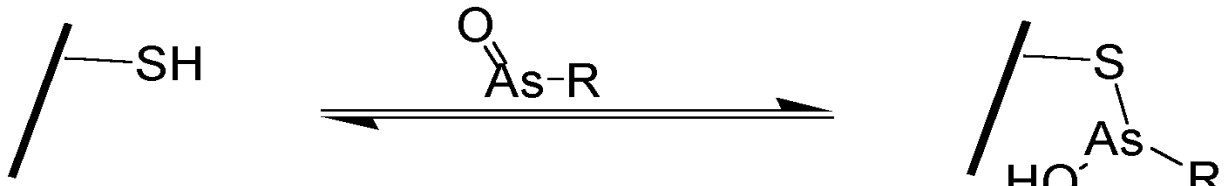


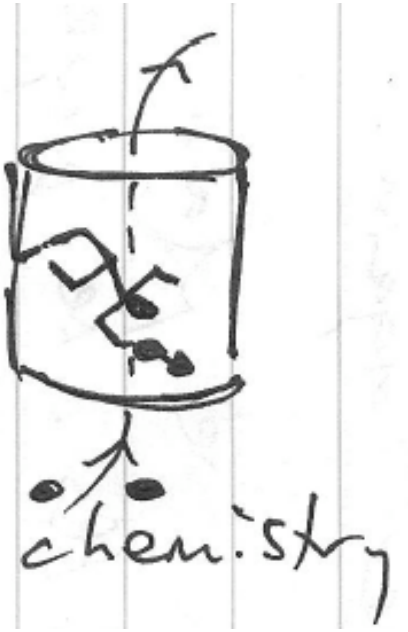


Ibuprofen


a
0
1
0
-20
b
0
2
0
-20
c
0
1
2
0
-20
3 ms
Level-0: βCD (trans)
Level-1: S-ibuprofen (cis)
Level-2: R-ibuprofen (cis)
Normalized Count / N
0.10
0.08
0.06
0.04
0.02
0.00
S
-8
-6
-4
-2
0
Amplitude / pA
R
S

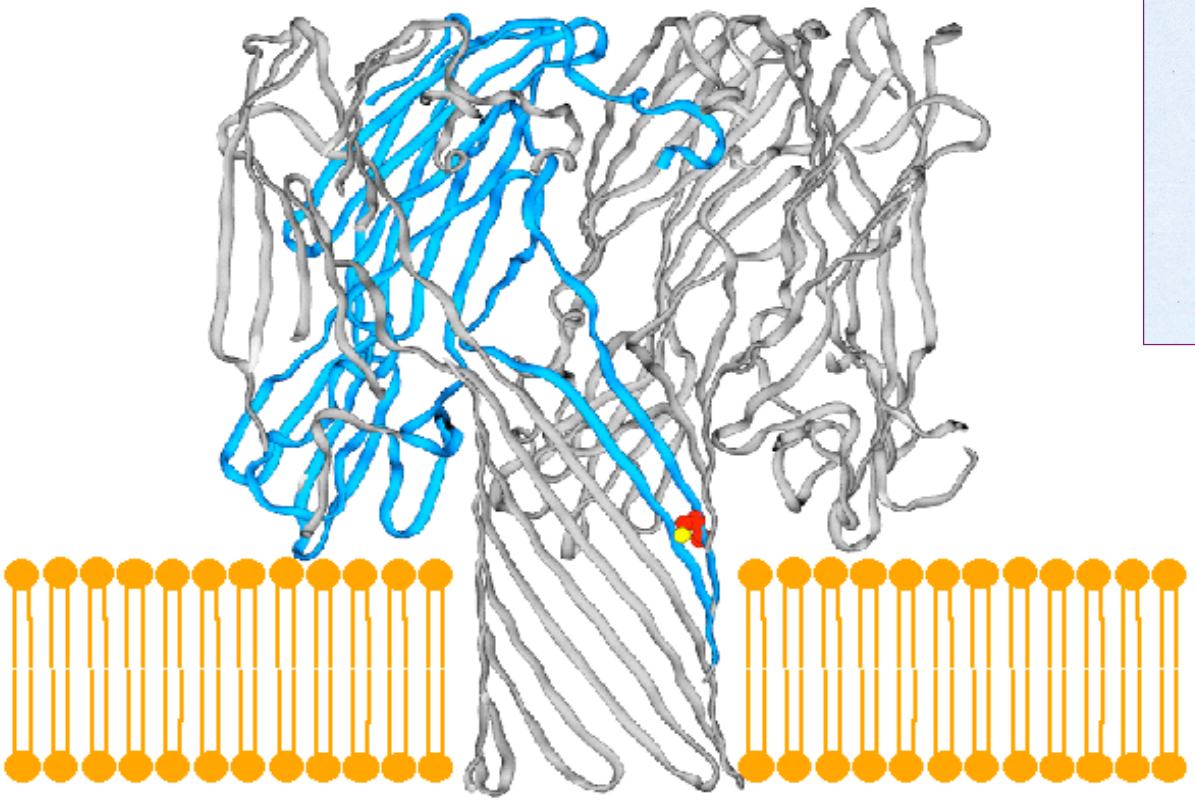
Single molecule covalent chemistry inside the pore

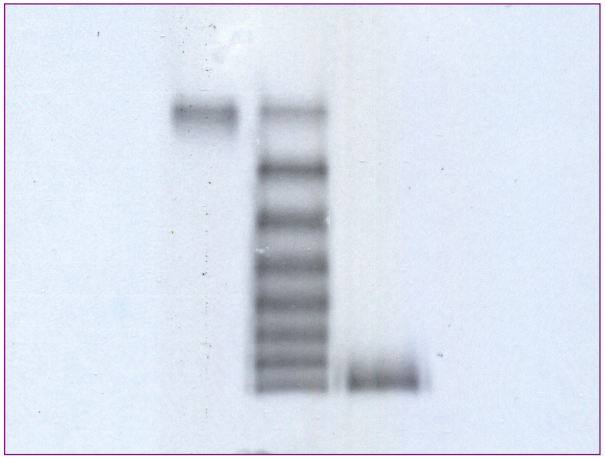




αHL pore containing a single cysteine

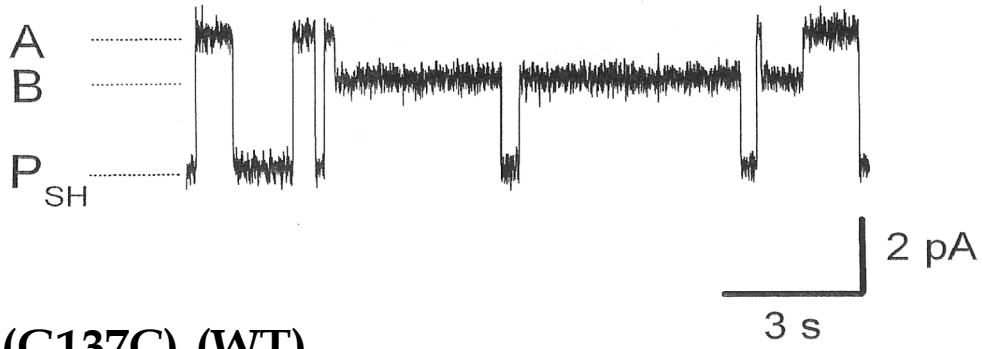
projecting into the channel lumen



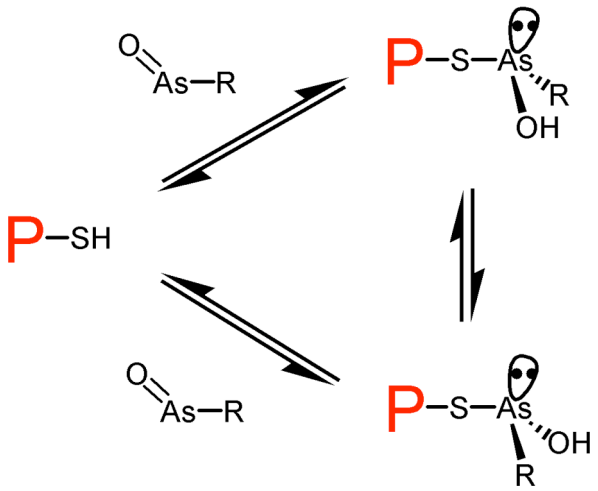


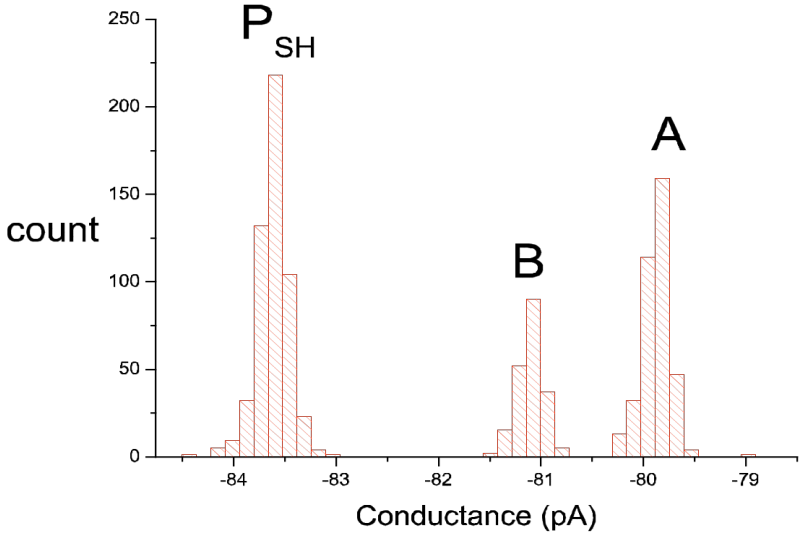
P_{SH}

Chiral arsenic adducts



$(G137C)_1(WT)_6$

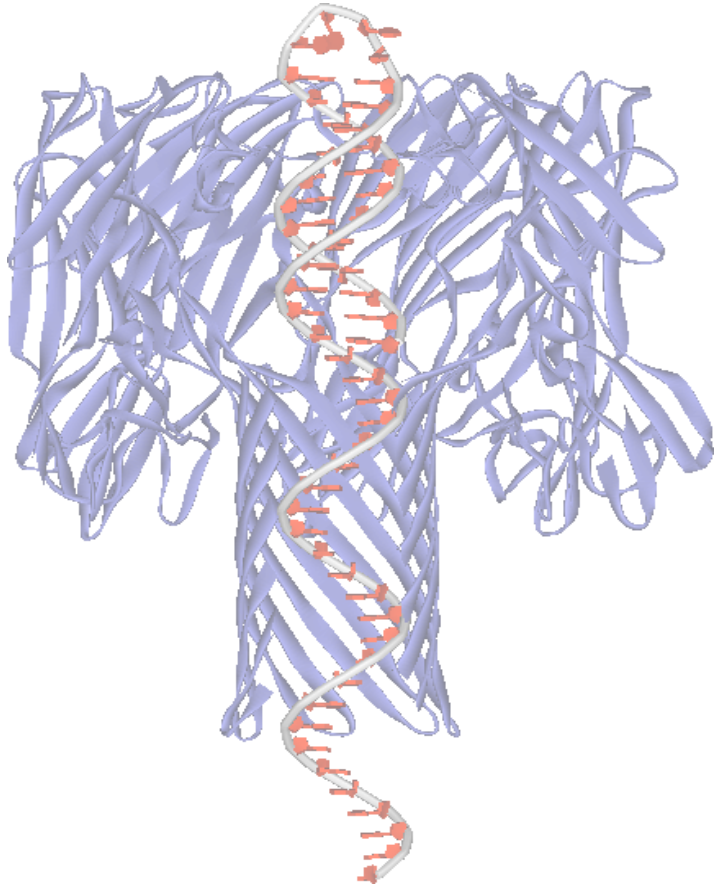




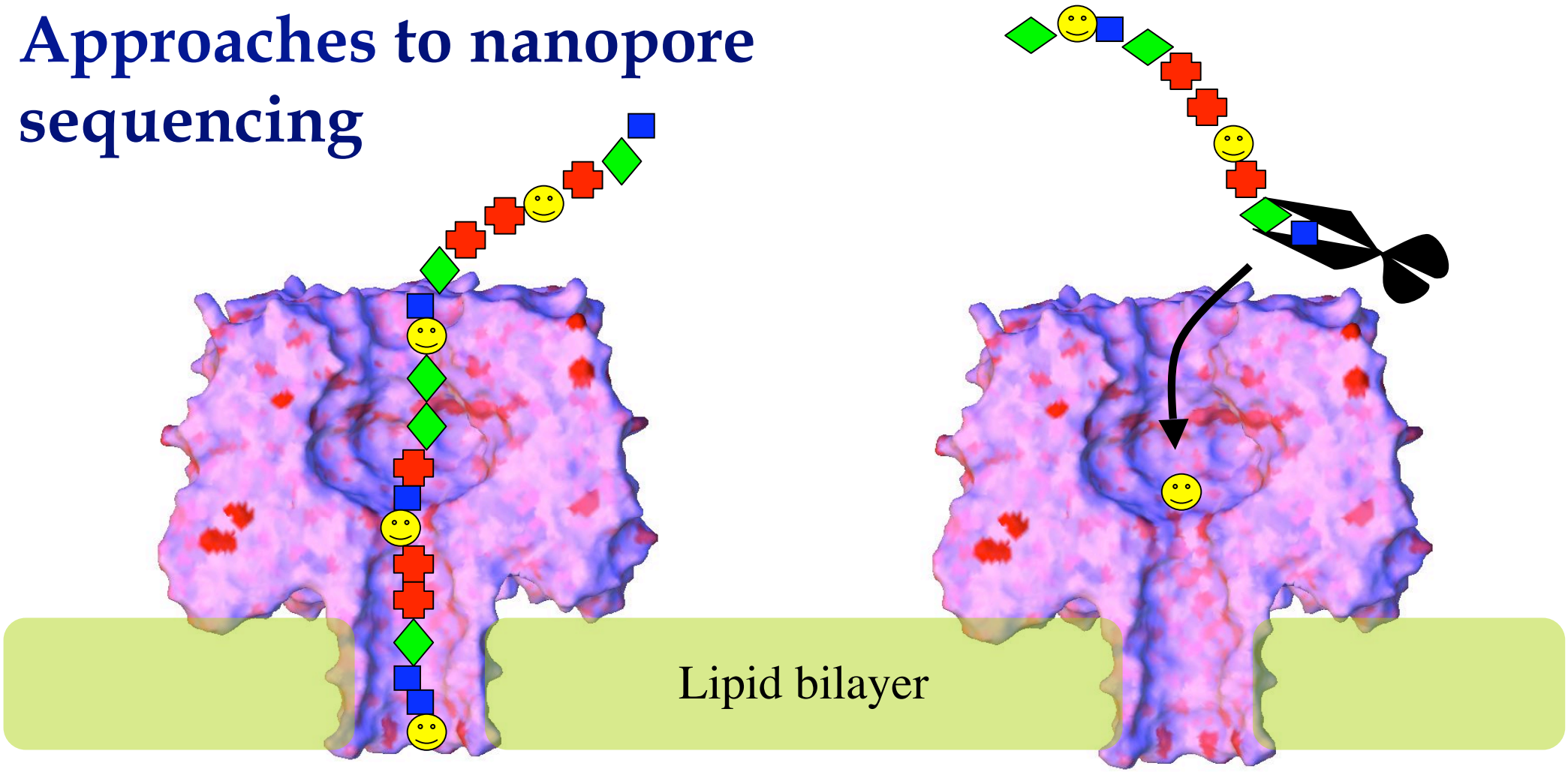
Towards single-molecule DNA sequencing with engineered nanopores

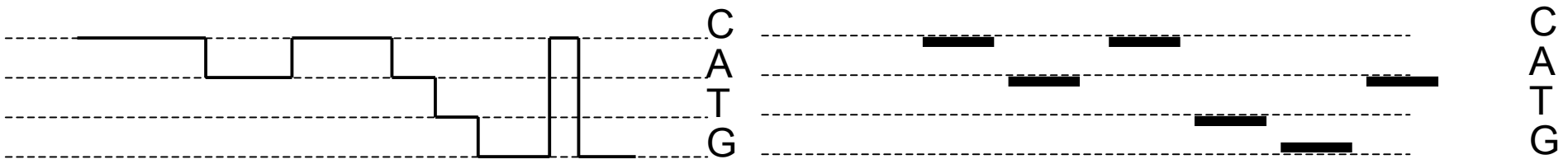
Hagan Bayley

M. Reza Ghadiri



Approaches to nanopore sequencing



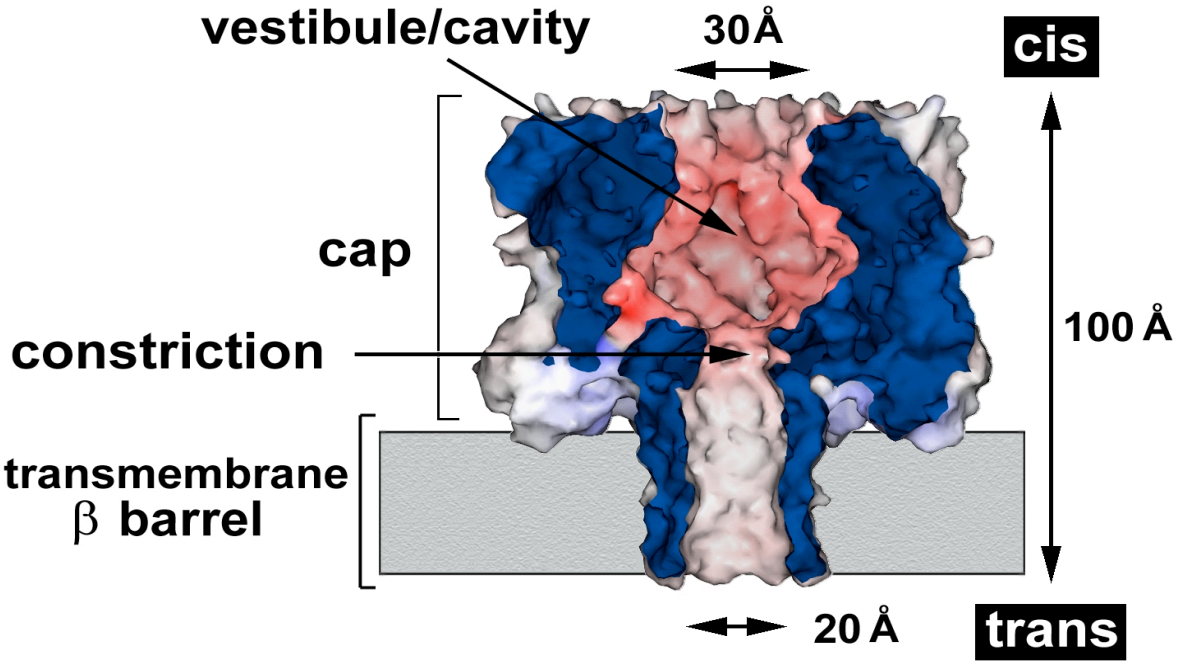


… unmodified DNA sequencing …

Nanopore DNA sequencing concept

- US 5,795,782: G. Church, D.W. Deamer, D. Branton, R. Baldarelli, J. Kasianowicz, issued 1998
- J. J. Kasianowicz *et al. PNAS*, **93**, 13770 (1996)

Transient transport studies:
Akeson, Marziali, Meller …



Discrimination between homopurine versus homopyrimidine

Central issues:

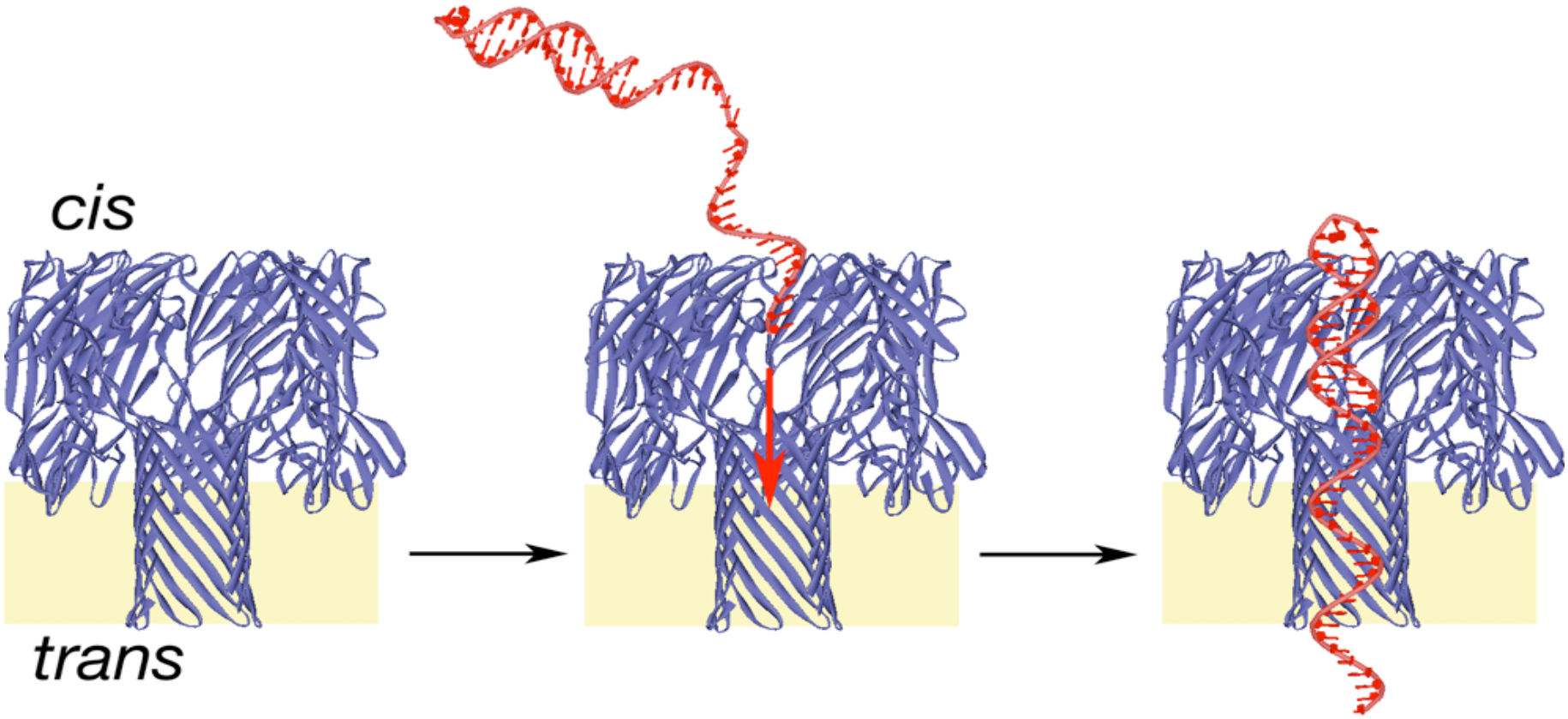
Molecular recognition

- Single nucleobase resolution
- Nucleobase discrimination (GATC)

DNA translocation speed

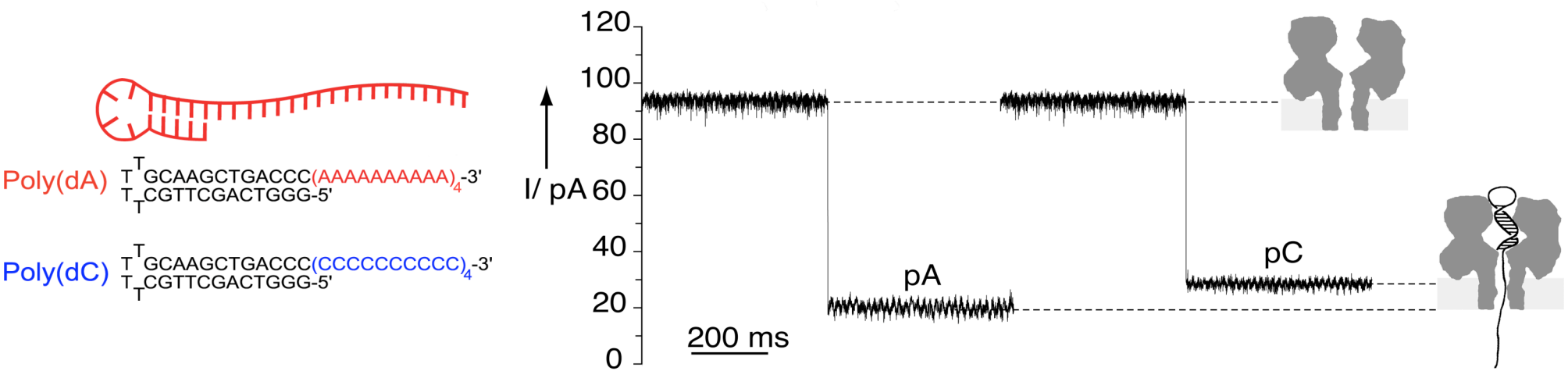
Key is the use of engineered pores …..

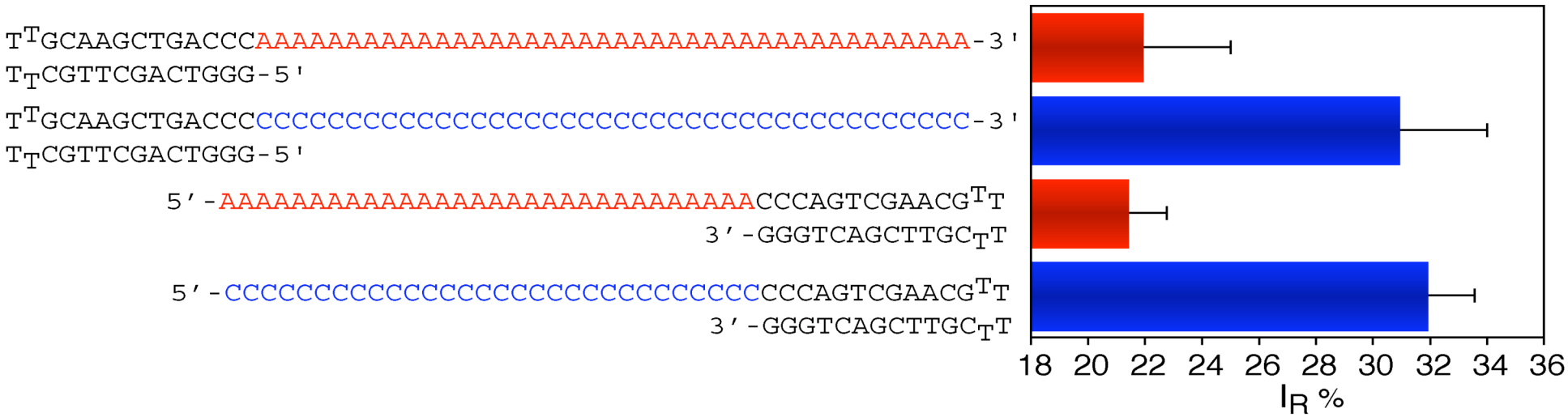
Probing nucleobase recognition capacity using single-species αHL•DNA pseudorotaxanes



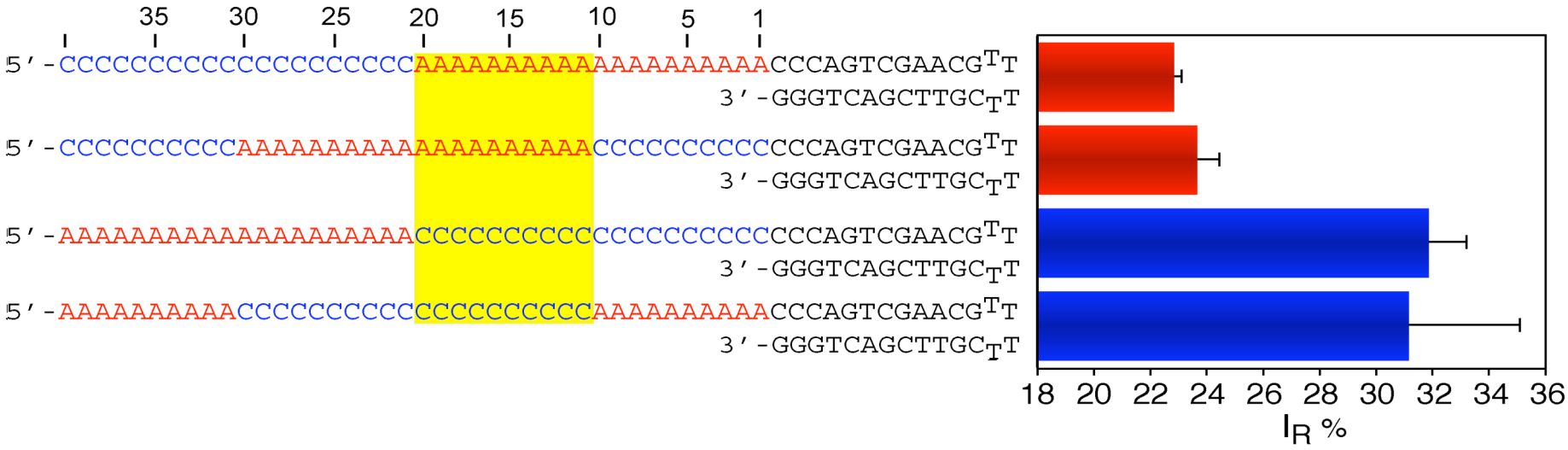
The ss-DNA can thread into the pore under positive applied potentials and be held stably in a pseudorotaxane configuration by the interactions of the bulky terminal hairpin segment at the pore entrance.

αHL•DNA pseudorotaxanes display characteristic ion conductance blockades for homopurine and homopyrimidine DNA strands

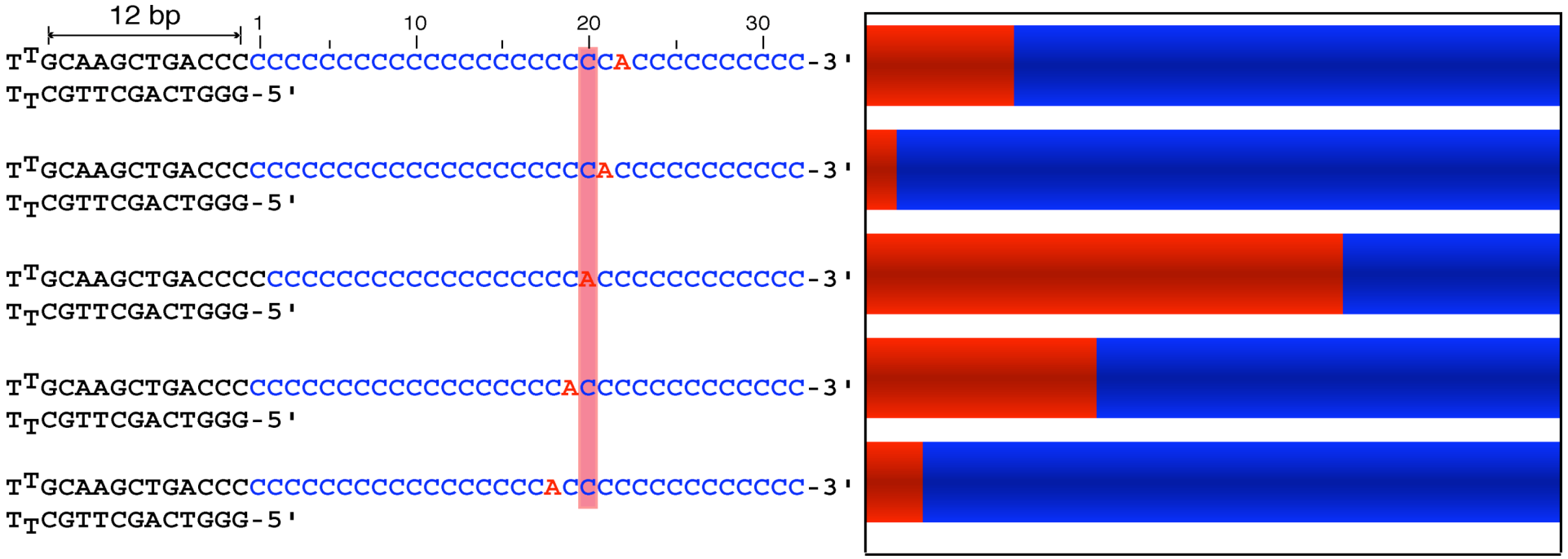




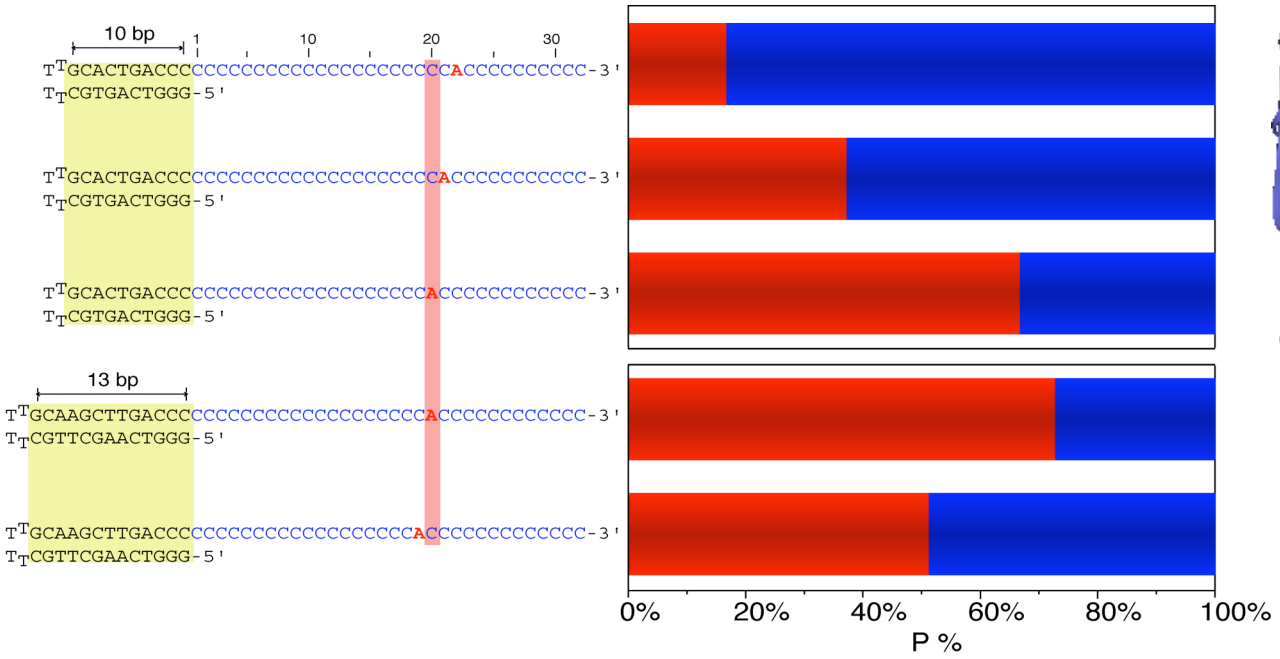
Probing the nucleobase spatial resolution

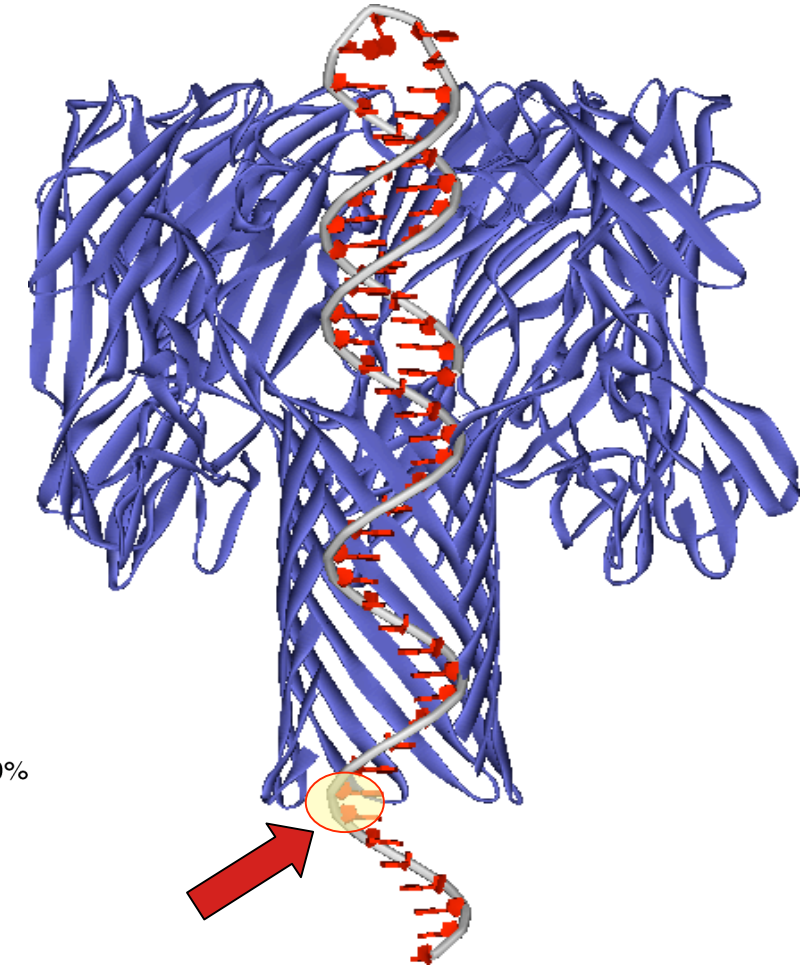


Native αHL has an apparent single nucleobase recognition capacity

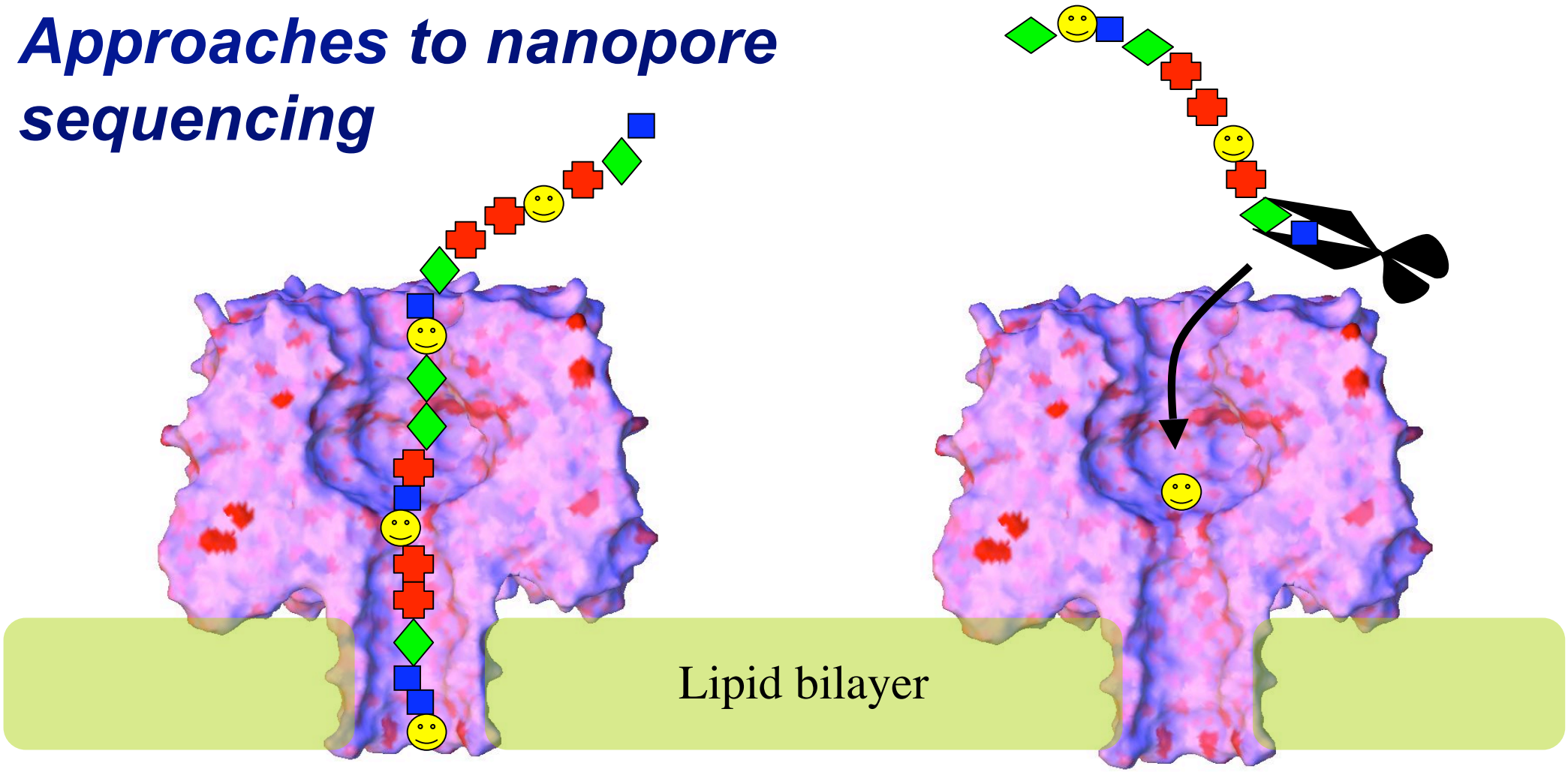


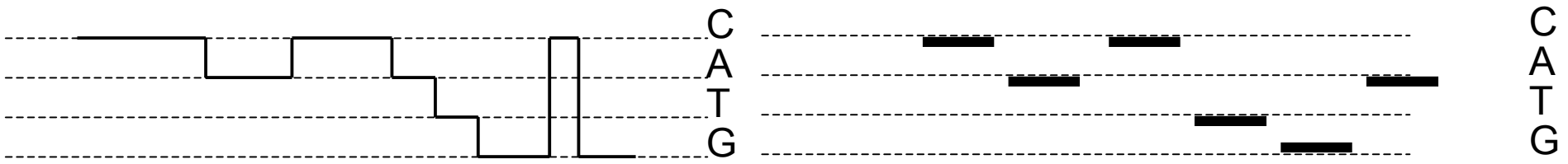
Single nucleotide recognition is independent of the DNA hairpin stem size suggesting that the location of the nucleobase recognition site is near the αHL trans exit.





Approaches to nanopore sequencing

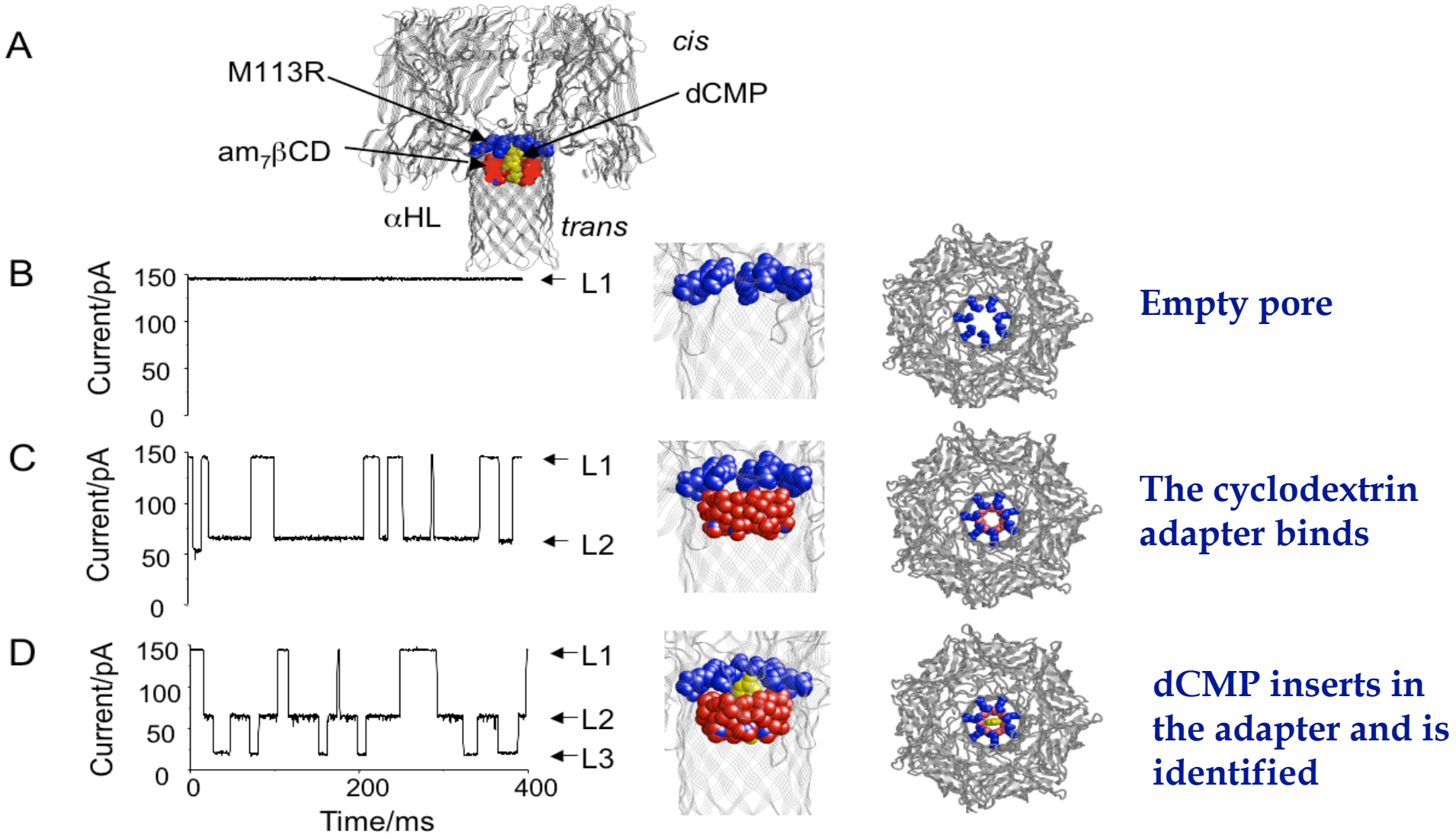




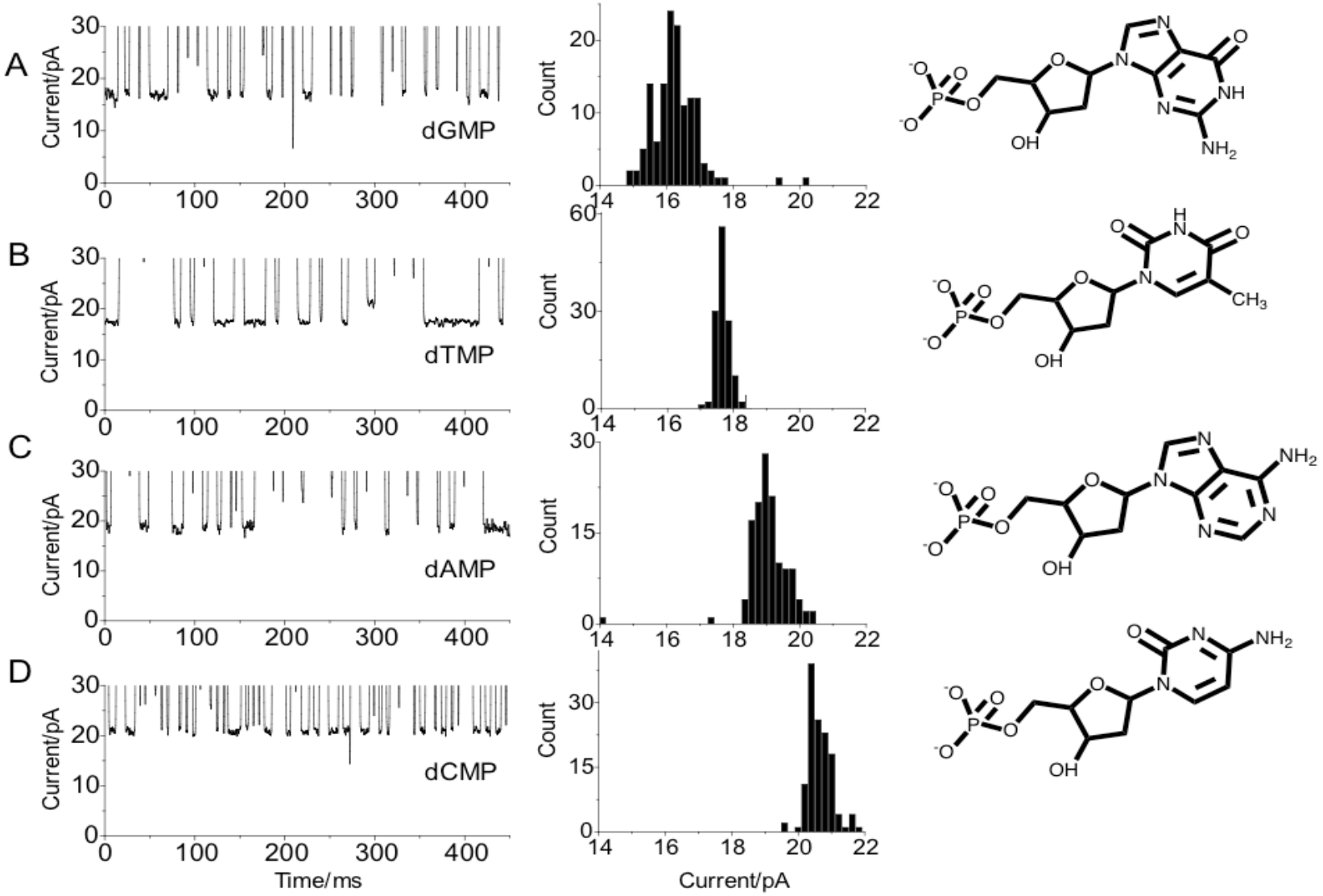
… unmodified DNA sequencing …

Recognition of nucleoside monophosphates

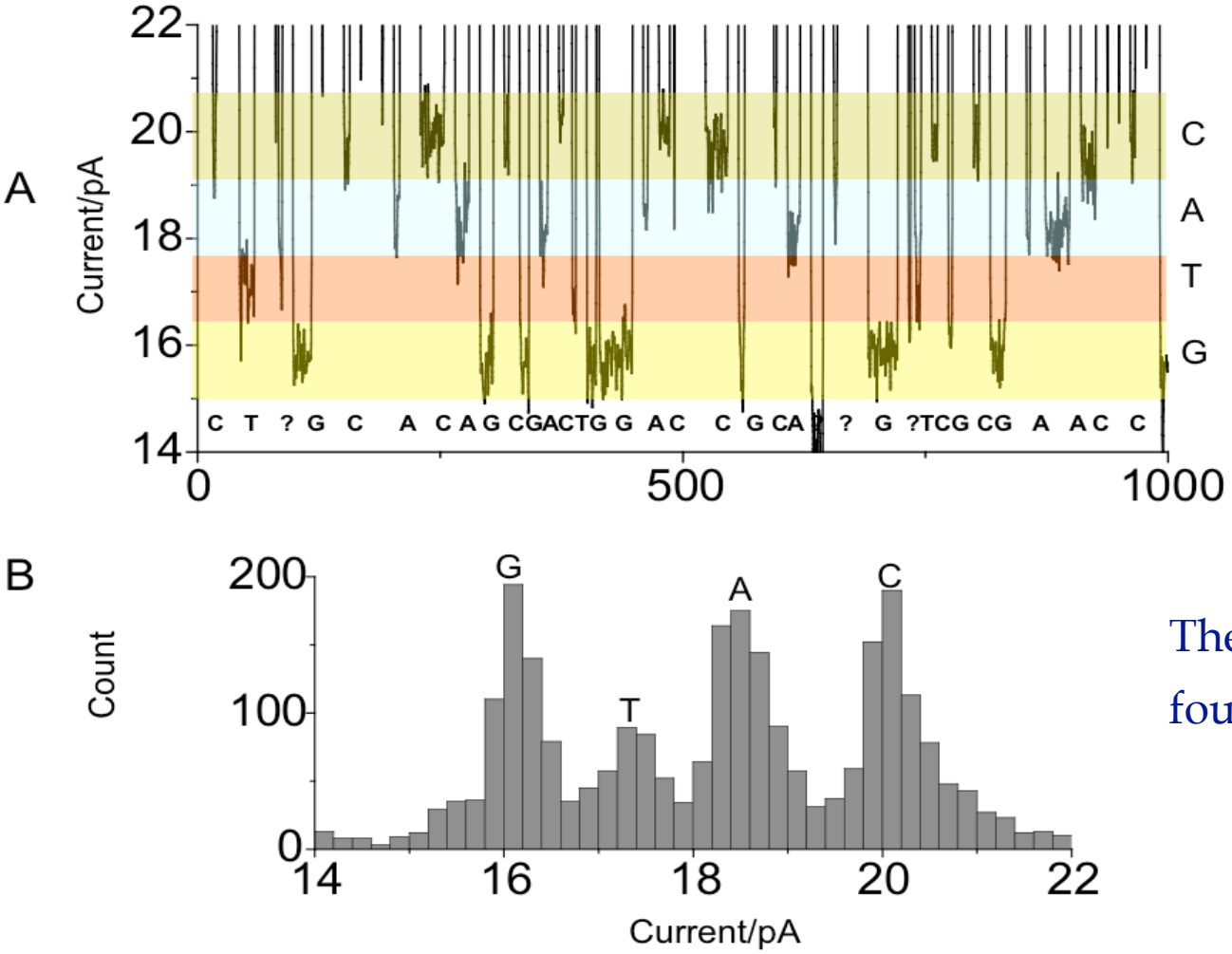
A molecular recognition site is formed by combining a genetically engineered arginine ring, and a cyclodextrin with seven amino groups in the primary positions



Each nucleoside 5′-monophosphate displays a different amplitude of current block allowing its identification



Different nucleoside 5′-monophosphates may be identified individually in a mix



The amplitude distribution shows four different populations

Summary & future goals:

Molecular recognition

- Single nucleobase resolution
- Nucleobase discrimination (GATC)

… our results demonstrate that resolution and discrimination can be achieved although improvements are needed

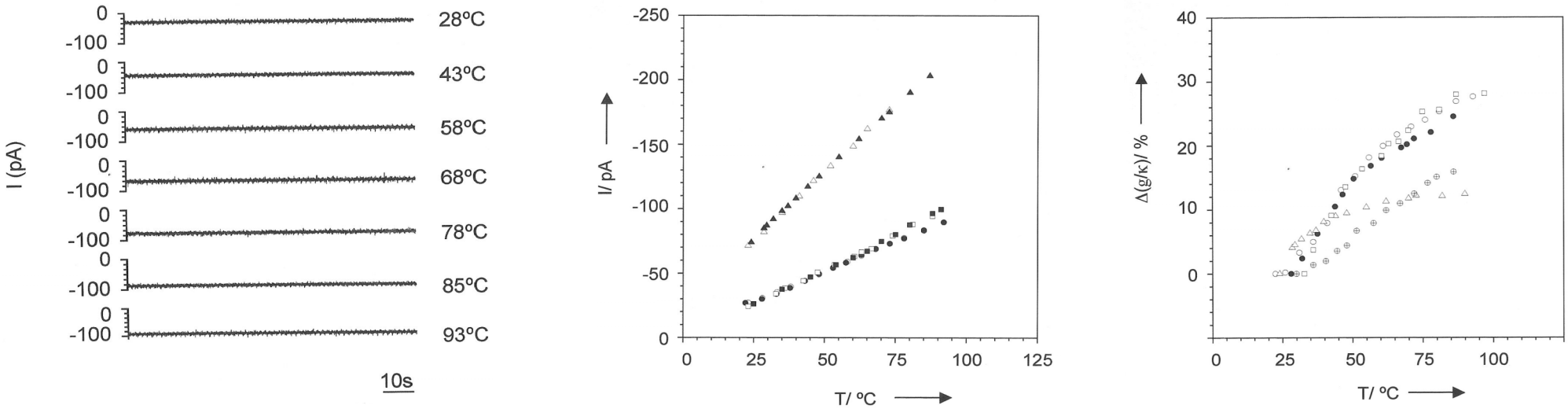
DNA translocation Speed

… use of engineered pores …

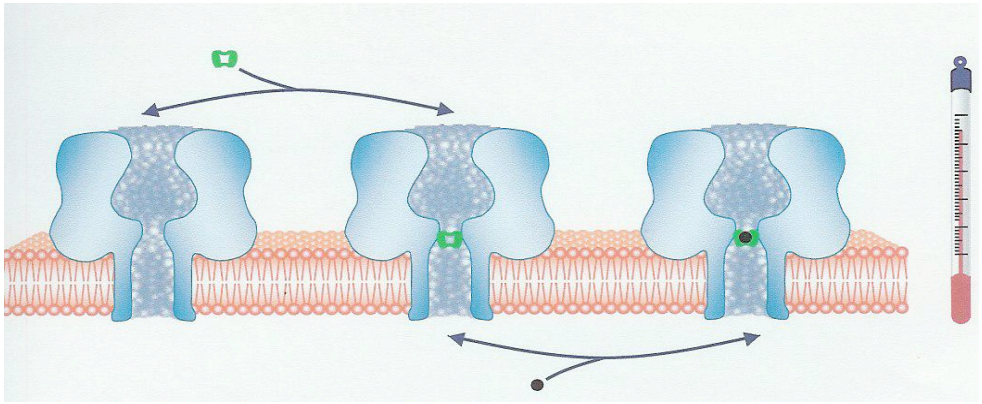
…molecular brakes, high internal viscosity, attached nucleic acid-handling enzymes, etc.

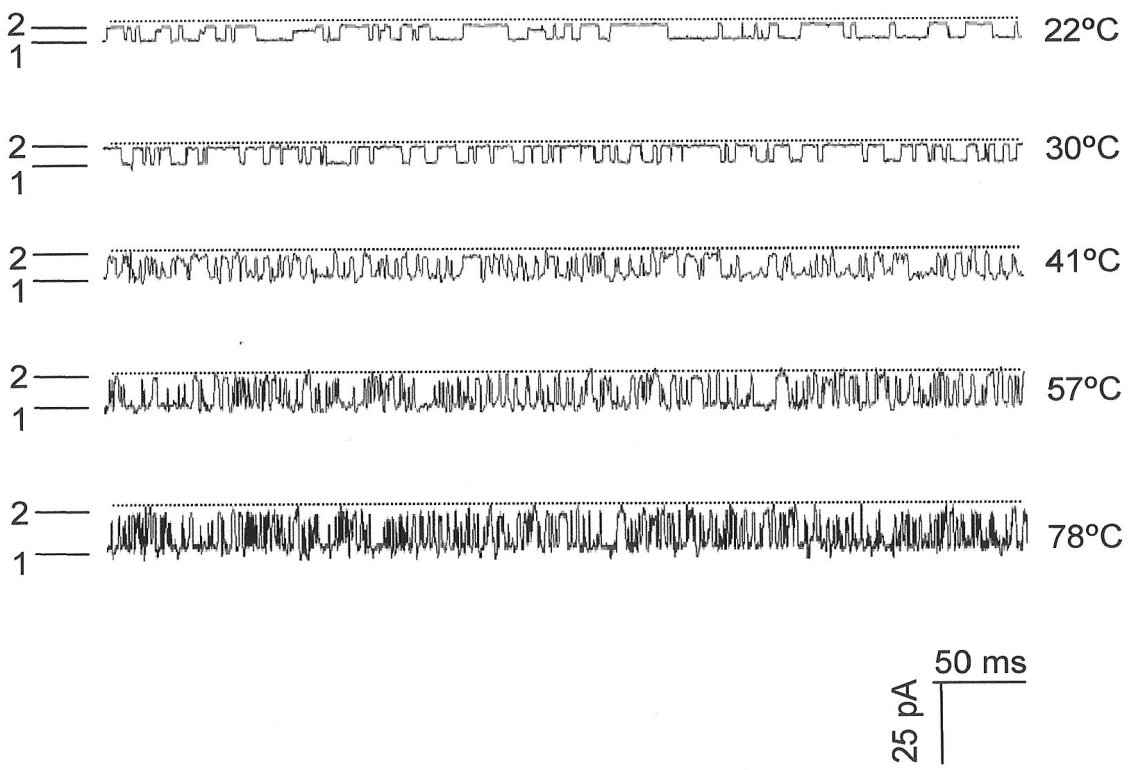
Protein pores at high temperatures

α-Hemolysin, leukocidin and OmpG are stable at high temperatures and show no sharp transitions in conductance

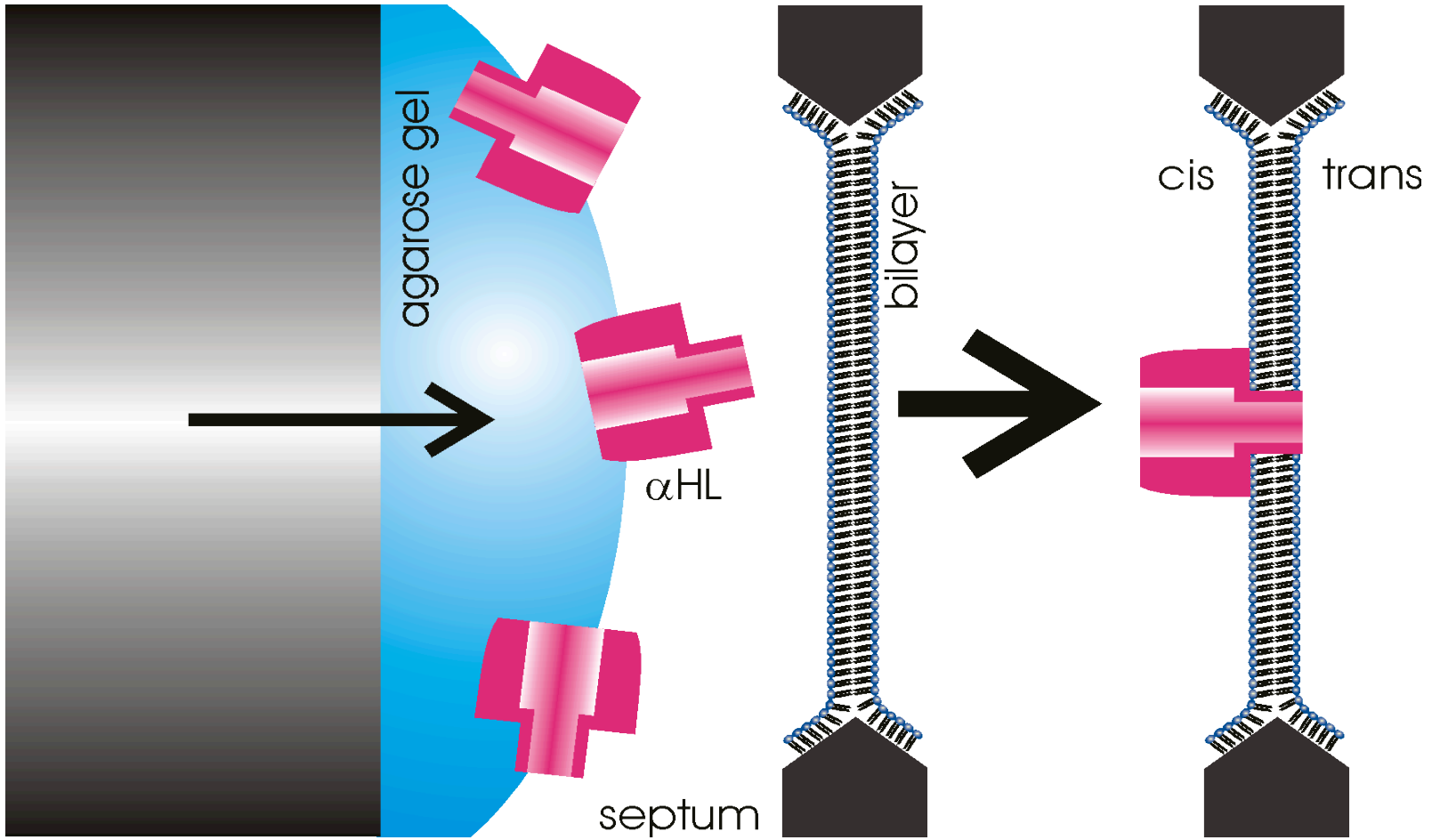


$(M113F/K147N)_7$, βCD, adamantane-1-carboxylate





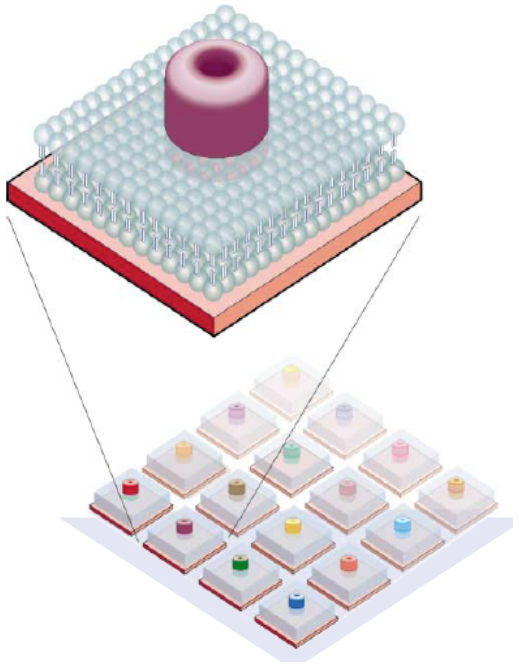
Direct insertion of single αHL pores using a mechanical probe



Holden, M.A. & Bayley, H. *J. Am. Chem. Soc.* **127**, 6502-6503 (2005)

Array fabrication might be facilitated by probe technology

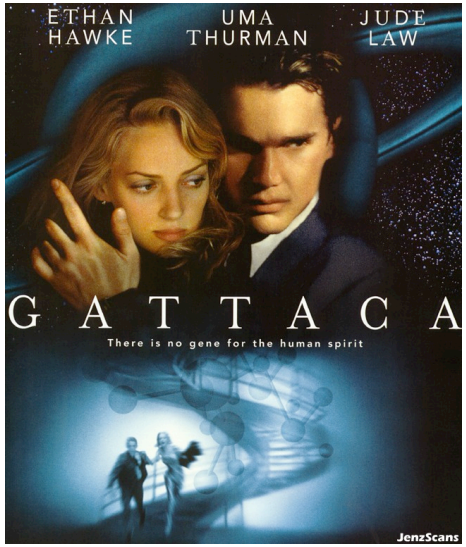
- Sensor arrays for stochastic sensing
- Rapid screening of membrane proteins
- DNA sequencing



- **Making a miniaturized & rugged device**

www.oxfordnanolabs.com

Oxford NanoLabs

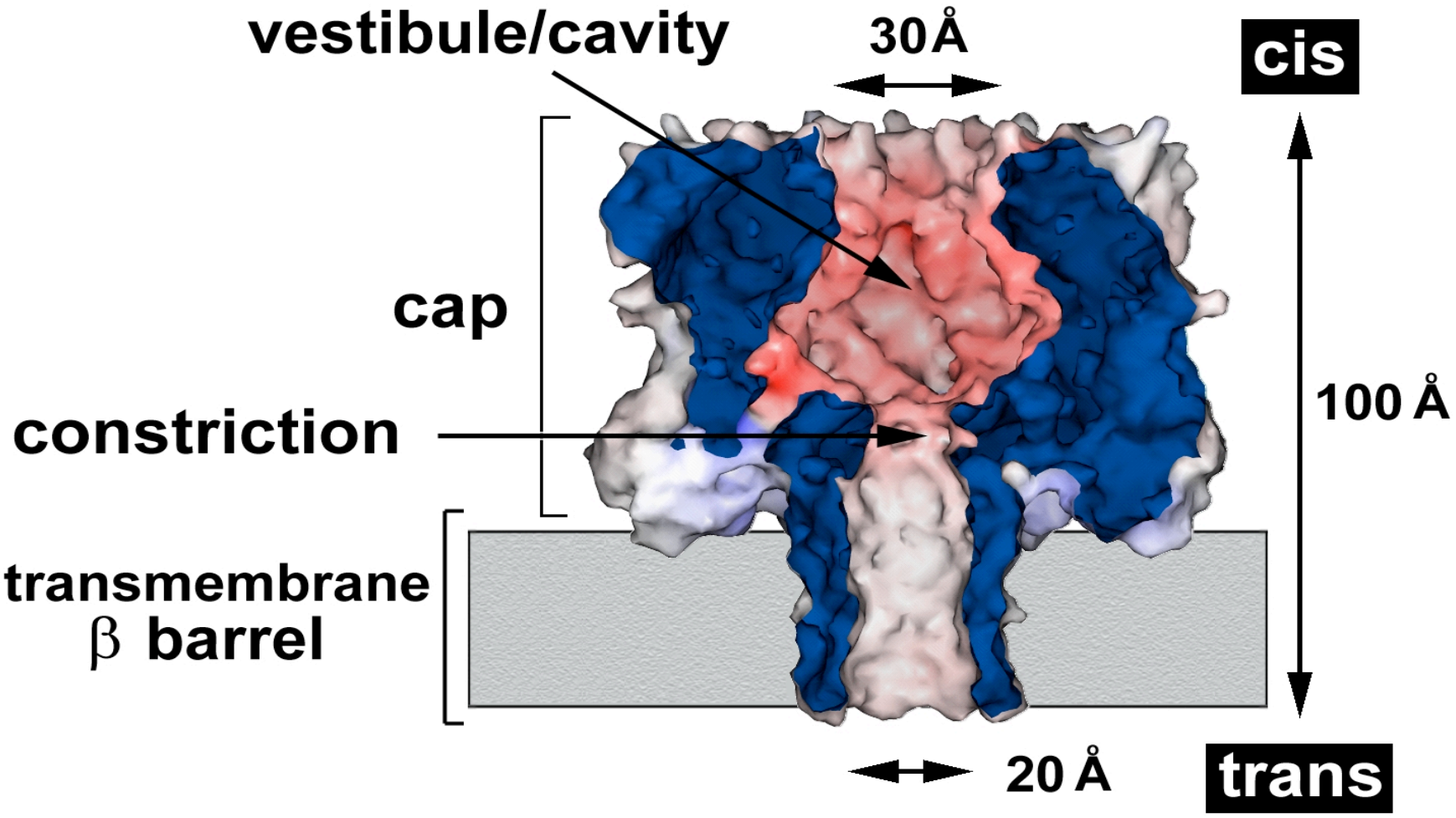


visit our web site:

www.chem.ox.ac.uk/bayleygroup/

*Xiaofeng Kang (Texas A&M University)
*Reza Ghadiri (Scripps)
*Nurit Ashkenasy (Scripps) ➔ U Negev

Oxford:
*Yann Astier
Arijit Banerjee
*Orit Braha
Min Chen
James Clarke
Oliver Daltrop
Anne Hammerstein
Andrew Heron
*Matt Holden
Lakmal Jayasinghe
Beatrice Li
Wen-wu Li
Siran Lu
Giovanni Maglia
Amy Mason
Rethi Madathil
Ellina Mikhailova
Dvir Rotem
*Seong-Ho Shin ➔ UC Berkeley
Mackay Steffenson
Haichen Wu


vestibule/cavity
30Å
cis
cap
constriction
100 Å
transmembrane
β barrel
20 Å
trans