PolyBayes

Ruslan Salakhutdinov

What is SNP

Source: A Science Primer.

- A Single Nucleotide Polymorphism, or SNP is a small genetic change, or variation, that can occur within a person's DNA sequence.

- An example of a SNP is the alteration of the DNA segment AAGGTTA to ATGGTTA

- Most SNPs are found outside of ”coding sequences”.

- SNPs found within a coding sequence are of particular interest to researchers because they are more likely to alter the biological function of a protein.

SNPs and Disease Diagnosis

Source: A Science Primer.

- Each person's genetic material contains a unique SNP pattern that is made up of many different genetic variations.

- Researchers have found that most SNPs are not responsible for a disease state.

- Instead, they serve as biological markers for pinpointing a disease on the human genome map:
  - Reason: they are usually located near a gene found to be associated with a certain disease.

- Occasionally, a SNP may actually cause a disease and, therefore, can be used to search for and isolate the disease-causing gene.

- We will see how a Bayesian method (PolyBayes) can be used to detect SNPs.

Bayes’ Rule

- For any hypothesis h and data d we have:

$$p(h|d) = \frac{p(d|h)p(h)}{\sum_{h \in H} p(d|h)p(h)}$$

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{P(data)}$$

- Idea. Suppose we have aligned DNA sequences (EST’s) of 10 individuals and we are looking at one specific position. We have two hypothesises: h_1 - there is a SNP or h_0 – there is no SNP.

- Well, the prior $p(h_1)$=0.003 and $p(h_0)$=1-0.003, because we believe that SNPs typically occur once every 333 bp.

- Once we observe the data (aligned ESTs) we can judge what are the posterior odds in favor of h_1.

Bayesian inference

Source: Josh Tenenbaum's example

- Data: John is coughing.
- Some hypotheses:

  1. John has a cold
  2. John has lung cancer
  3. John has a stomach flu

- Prior $P(h)$ favors 1 and 3 over 2
- Likelihood $P(d|h)$ favors 1 and 2 over 3
- Posterior $P(h|d)$ favors 1 over 2 and 3

What they do in the paper: PolyBayes

- The goal of the paper is to find SNPs from ESTs, pieces of DNA sequence, of 10 genomic clones (of 10 individuals).

- ESTs are small pieces of DNA sequence (usually 200 to 500 nucleotides long) that are generated by sequencing either one or both ends of an expressed gene.

- How they do it:
  - First obtain ESTs and construct an alignment against a fragment of the finished human reference sequence (less than 1 error per 10.000 bp). Draw this on the board.

  - Identify paralogues. These are the sequences that represent highly similar regions duplicated elsewhere in the genome. They may give rise to false SNP predications.

  - Use multiple alignment of sequences to detect SNPs using PolyBayes.

Identifying paralogues

- Is the number of mismatches observed between the genomic reference sequence and a matching EST was consistent with polymorphic variation as opposed to sequence difference between duplicated chromosomal locations.
- Key observation: Most ”paralogous” sequences exhibit a pair-wise dissimilarity rate higher than P_{PAR} = 0.02 (2%).
- This is compared with the average pair-wise polymorphism rate, P_{POLY} = 0.001 (0.1%).
- So, in a pair-wise match of length L, we'd expect LP_{POLY} mismatches due to polymorphism, versus LP_{PAR} mismatches due to paralogous difference.
- We also add E of mismatches that are expected to arise from sequencing errors.

Overall picture

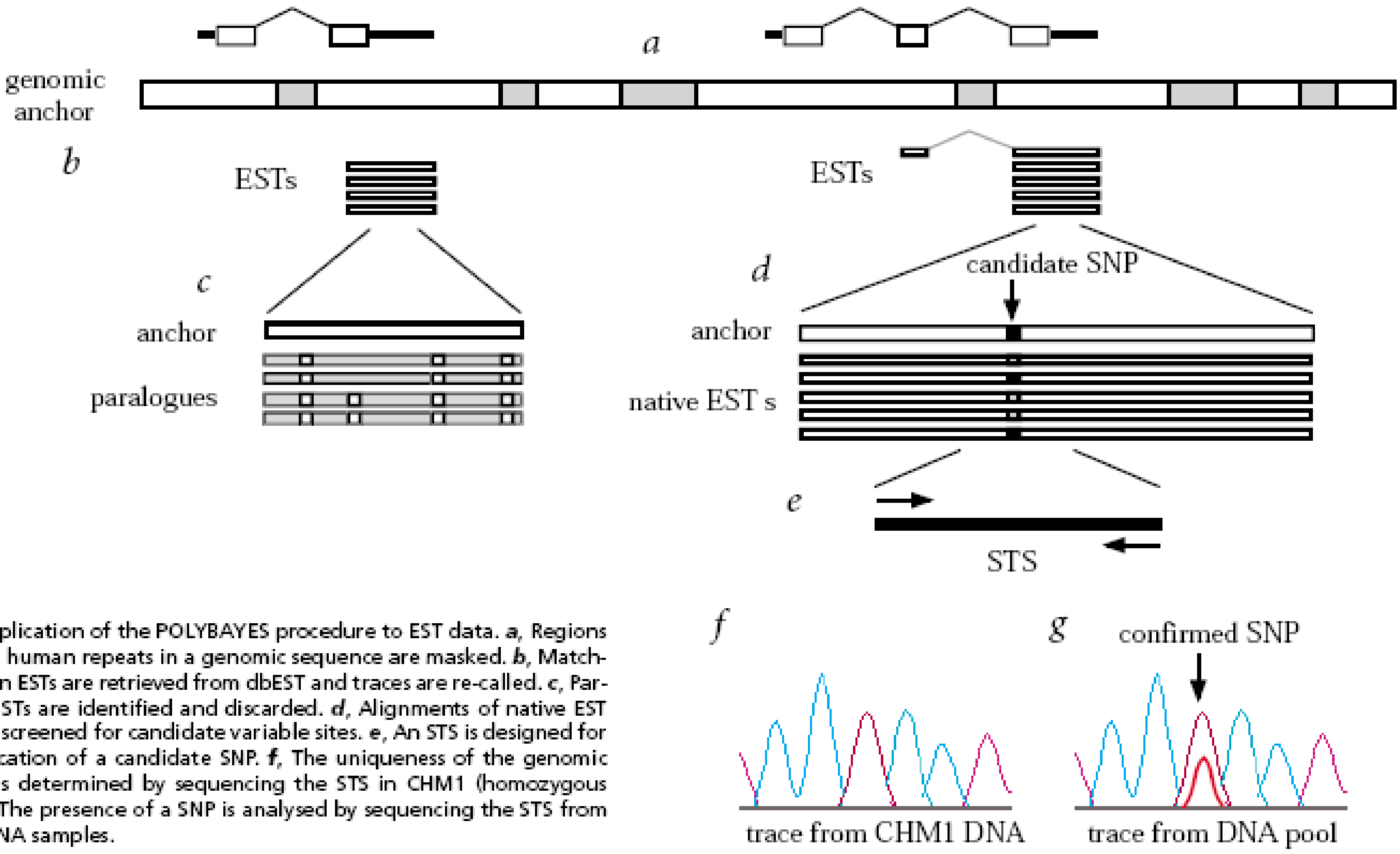


Fig. 1 Application of the POLYBAYES procedure to EST data. ***a***, Regions of known human repeats in a genomic sequence are masked. ***b***, Matching human ESTs are retrieved from dbEST and traces are re-called. ***c***, Paralogous ESTs are identified and discarded. ***d***, Alignments of native EST reads are screened for candidate variable sites. ***e***, An STS is designed for the verification of a candidate SNP. ***f***, The uniqueness of the genomic location is determined by sequencing the STS in CHM1 (homozygous DNA). ***g***, The presence of a SNP is analysed by sequencing the STS from pooled DNA samples.

Identifying paralogues: The model

- We have two models: M_{NAT} and M_{PAR}.

- The probability (the likelihood) of observing d discrepancies is approximated by the Poisson distributions with parameters:
  - $\lambda = D_{NAT} = LP_{POLY} + E$ for model M_{NAT}
  - $\lambda = D_{PAR} = LP_{PAR} + E$ for model M_{PAR}.

  Remember the Poisson:

$$p(D = d|\lambda) = e^{-\lambda}\frac{\lambda^d}{d!}$$

- Now, since we don't have any preference for either model, we use uninformative prior, or $P(M_{NAT}) = p(M_{PAR}) = 0.5$.

- Crank up Bayesian inference to get:

$$p(M_{NAT}|d) = \frac{p(d|M_{NAT})p(M_{NAT})}{p(d|M_{NAT})p(M_{NAT}) + p(d|M_{PAR})p(M_{PAR})}$$

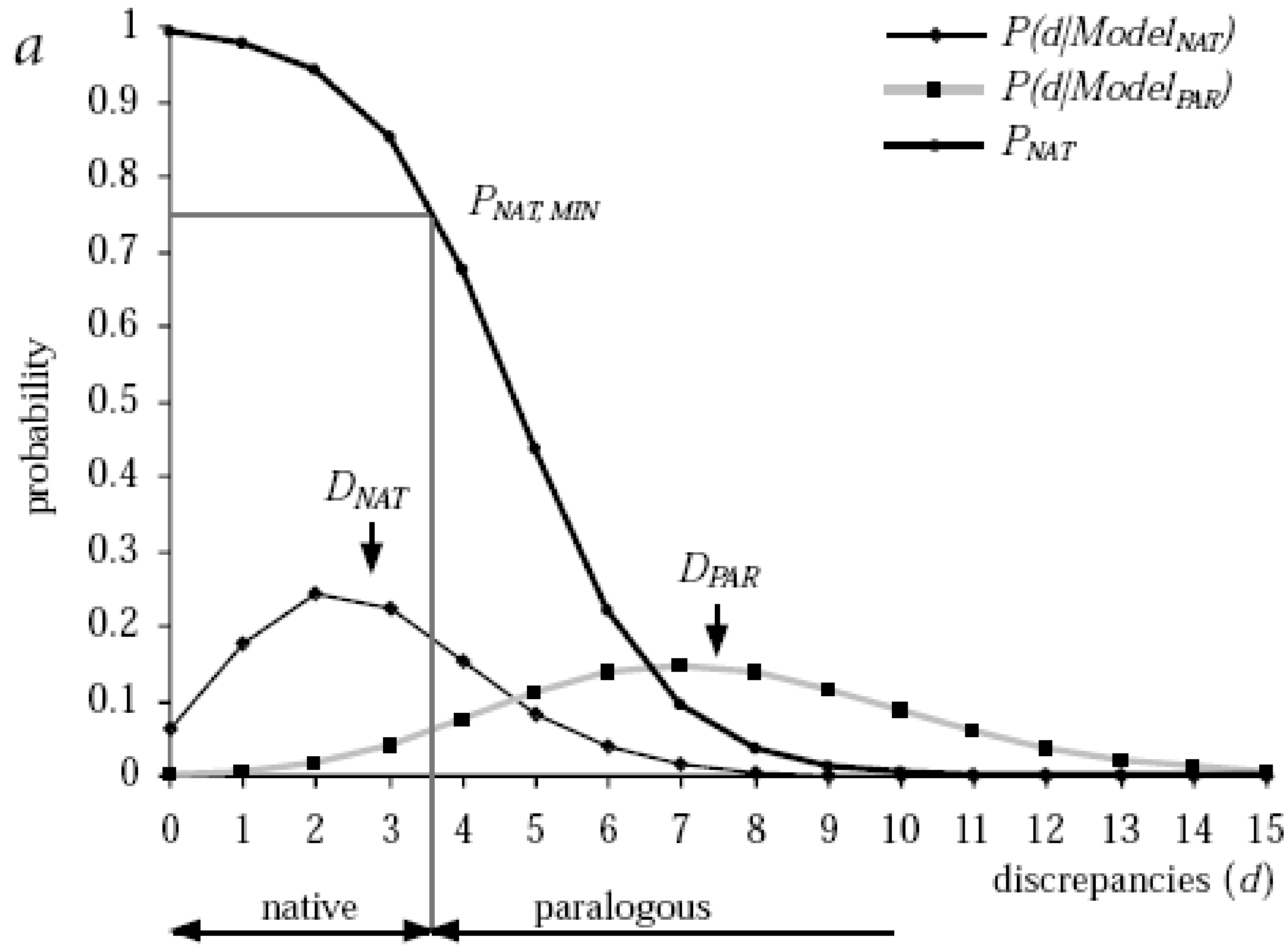
Bayesian Inference

- Crank up Bayesian inference to get:

$$p(M_{NAT}|d) = \frac{p(d|M_{NAT})p(M_{NAT})}{p(d|M_{NAT})p(M_{NAT}) + p(d|M_{PAR})p(M_{PAR})}$$

$$p(M_{NAT}|d) = \frac{1}{1 + e^{D_{NAT}-D_{PAR}}\left(D_{PAR}/D_{NAT}\right)^{d}}$$

Bayesian Inference



SNP detection in multiple alignments

- Suppose we have N cross-sections of a multiple alignment, R_1,,R_N. (Draw on the board).

- We want to identify polymorphic (as opposed to monomorphic) locations by evaluating the likelihood of nucleotide heterogeneity within cross-sections of a multiple alignment.

- Each of the nucleotides, S_1,...,S_N, in a cross-section of N sequences, can be any one of the four DNA bases, for a total of 4N nucleotide permutations.

- The likelihood, $P(S_i|R_i)$=1-P_{err} for the called base and $P(S_i|R_i)$=$P_{err}/3$ for each of the three uncalled bases.

SNP detection in multiple alignments

- Total a priori probability that a site is polymorphic is $P_{poly} = 0.003$.

- So the values P_{poly} have to be distributed to assign a prior probability $P(S_1, ..., S_N)$ to each polymorphic permutation.

- $(1 - P_{poly})/4$ is assigned to each of the four non-polymorphic permutations, corresponding to a uniform base composition, $P(S_i)$.

- What the heck does that mean? Show an example.

Bayesian Inference

- Once we have defined our likelihoods and priors, we can estimate the posterior probabilities of a particular permutation:

$$p(S_1, S_2|R_1, R_2) = \frac{p(R_1, R_2|S_1, S_2)p(S_1, S_2)}{p(R_1, R_2)} =$$

$$\frac{p(R_1|S_1)p(R_2|S_2)}{p(R_1, R_2)}p(S_1, S_2) \sim p(R_1|S_1)p(R_2|S_2)p(S_1, S_2)$$

  Note that

$$p(R_1|S_1) = \frac{p(S_1|R_1)p(R_1)}{p(S_1)}$$

  Thus

$$p(S_1, S_2|R_1, R_2) \sim \frac{p(S_1|R_1)}{p(S_1)}\frac{p(S_2|R_2)}{p(S_2)}p(S_1, S_2)$$

- The Bayesian posterior probability of a SNP is the sum of posterior probabilities of all heterogeneous permutations observed in the cross section.

Bayesian Inference

- The Bayesian posterior probability of a SNP is the sum of posterior probabilities of all heterogeneous permutations observed in the cross section.

- Candidate SNP is identified if the corresponding SNP posterior probability exceeded a threshold value of 0.40.

- Bayesian model takes into account
  - depth of coverge (N)
  - base quality values of the sequences P_{err}
  - a priori expected rate of polymorphic sites in region (P_{poly}).

- And like all other papers they show fantasic results.

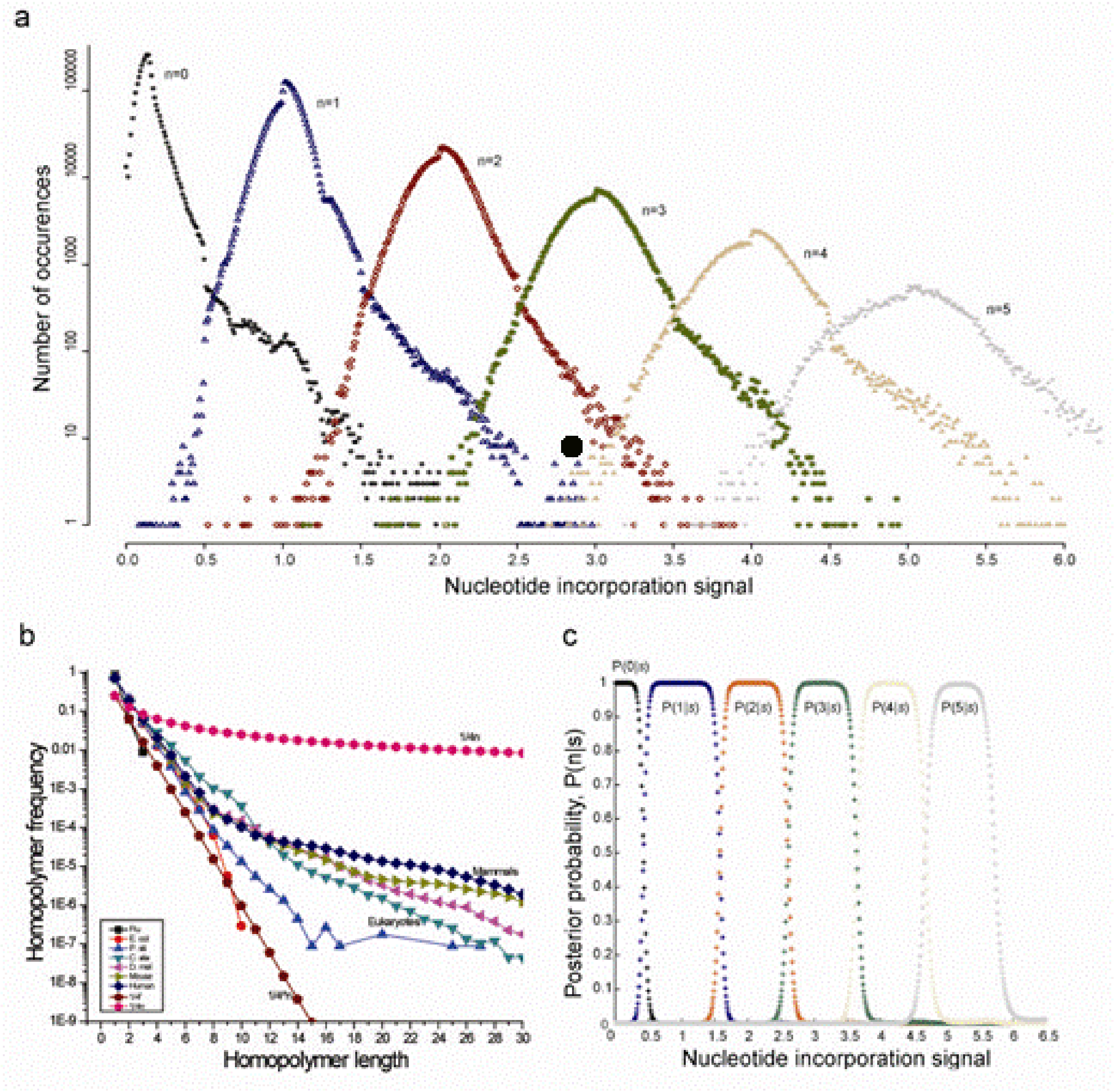
A bit on PyroBayes

- We have the sequencing reads produced by the 454 Life Sciences pyrosequencers.

- The light intensity signal observed in each cycle is proportional to the actual number of incorporated nucleotides.

- The signal for a fixed number of incorporated bases (e.g. a homopolymer AAA) varies substantially, and there is usually a nonzero signal even when no base is incorporated.

PyroBayes


a
Number of occurences
1
10
100
1000
10000
100000
n=0
n=1
n=2
n=3
n=4
n=5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
Nucleotide incorporation signal
b
Homopolymer frequency
1
0.1
0.01
1E-3
1E-4
1E-5
1E-6
1E-7
1E-8
1E-9
1/4n
Mammals
Eukaryotes
1/4^n
E. coli
D. mel
Mouse
Human
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
Homopolymer length
c
P(0|s)
P(1|s)
P(2|s)
P(3|s)
P(4|s)
P(5|s)
Posterior probability, P(n|s)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5
6
6.5
Nucleotide incorporation signal

A bit on PyroBayes

- Let s is the observed nucleotide incorporation signal and n is the homopolymer length.

- Use observed frequencies as estimates for the data probabilities $p(s|n)$.

- For the prior probability values $p(n)$, use the average frequency of the eukaryote homopolymer frequencies.

- Crank up Bayesian inference (up to n=100):

$$p(n|s) = \frac{p(s|n)p(n)}{\sum_{k=1}^{100} p(s|k)p(k)}$$

- The number n for which this posterior probability is highest is the most likely number of bases.

The End