

Factoring local sequence composition in motif significance analysis

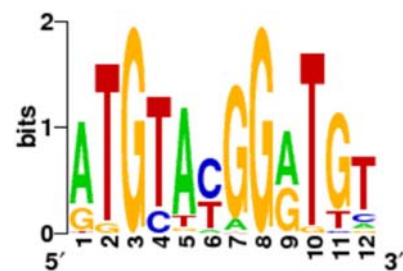
Patrick Ng and Uri Keich

Cornell University, Ithaca NY, USA

GIW 2008

Motif Finding

- DNA Motif – over-represented nucleotide **pattern** that has biological significance
- Why are we interested in finding motifs?
 - Transcription factor binding sites
- Over 100 motif-finders
 - MEME (Bailey *et al.* 1994)
 - Weeder (Pavesi *et al.* 2004)
 - YMF (Sinha *et al.* 2000)
 - BioProspector (Liu *et al.* 2001)



Motif Statistical Significance

- Does the motif I found have **biological** significance?
 - **Hard problem**
 - Require biological wet-lab experiment (e.g. knockout)
- Is the motif I found **statistically** significant? Is it special comparing to a motif I found in a random set of sequences?
 - **Easier problem**
 - Time and cost efficient
 - Ensemble algorithms
 - Combine results of different motif-finders
 - Combine results over different parameters of a finder

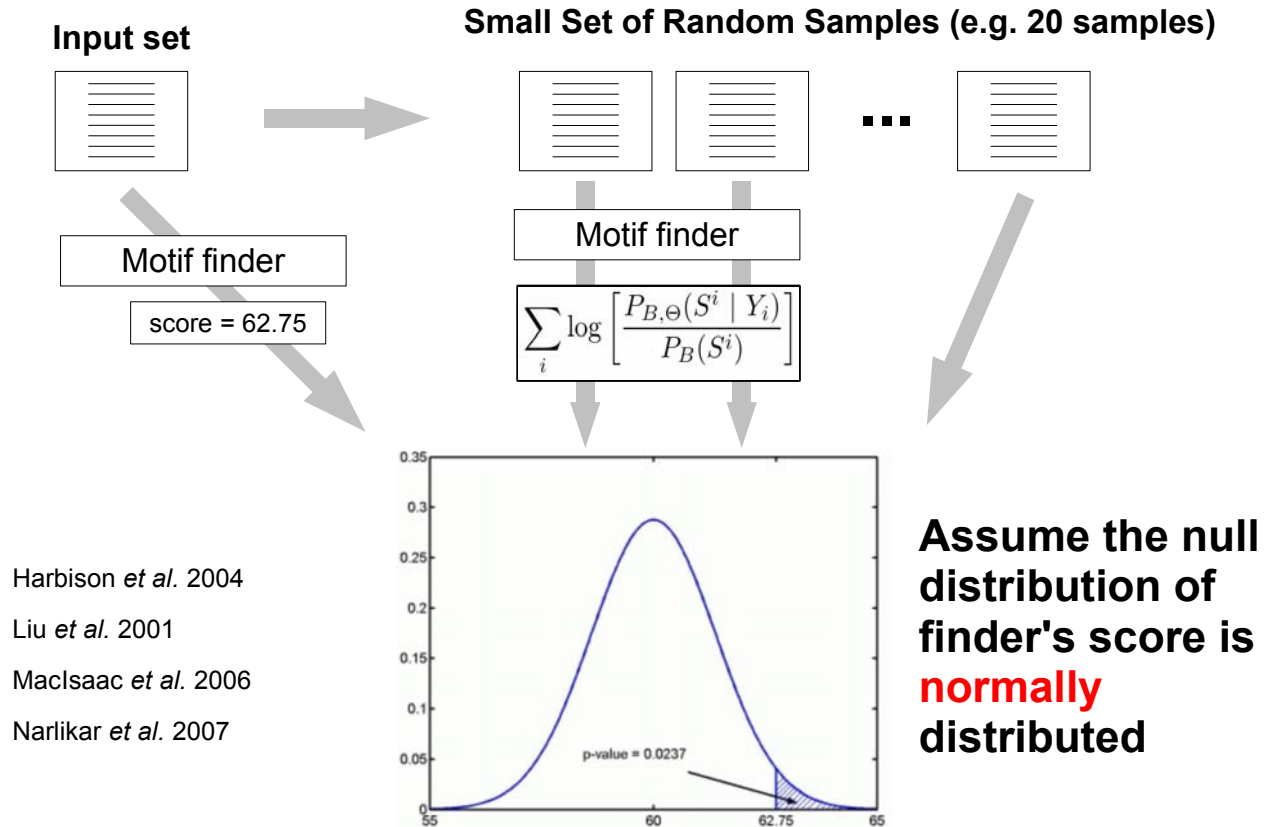
3

Issues and Challenges

- Motif-finders are **slow** (~30 sec. to 5 min.)
 - Naive Monte-Carlo estimate of p -value is intractable
 - e.g. In multiple testing, we may need p -value < 0.001
 - 30 seconds * 10,000 random sets \approx 3 days
- How does one define a **better** motif?
 - Different motif-finders use different scoring metrics
 - Likelihood ratio, Enrichment score, etc.
- What is a **random** set of DNA sequences?
 - Input data has different composition than a random chosen sequence (e.g. GC-content, dinucleotide frequency)

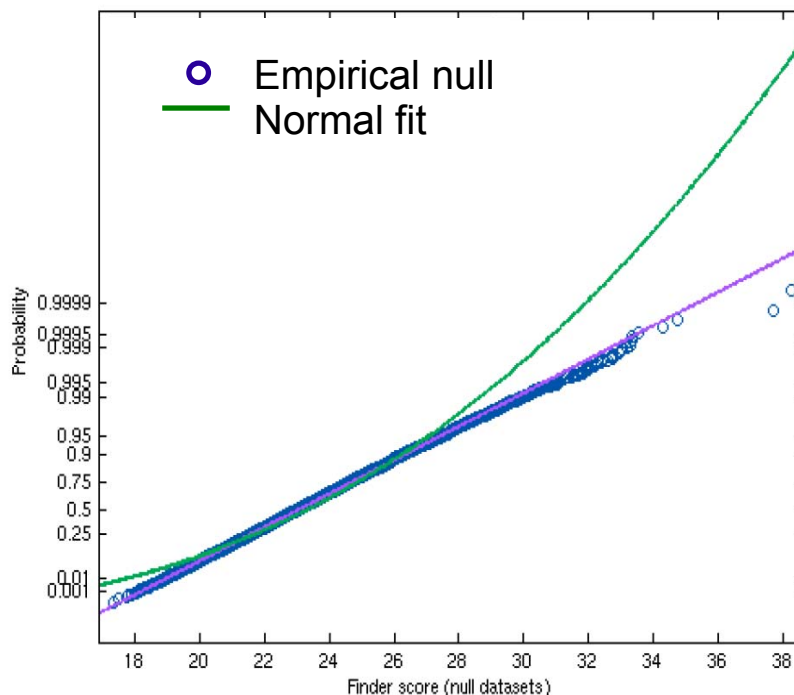
4

A Parametric Motif Significance



5

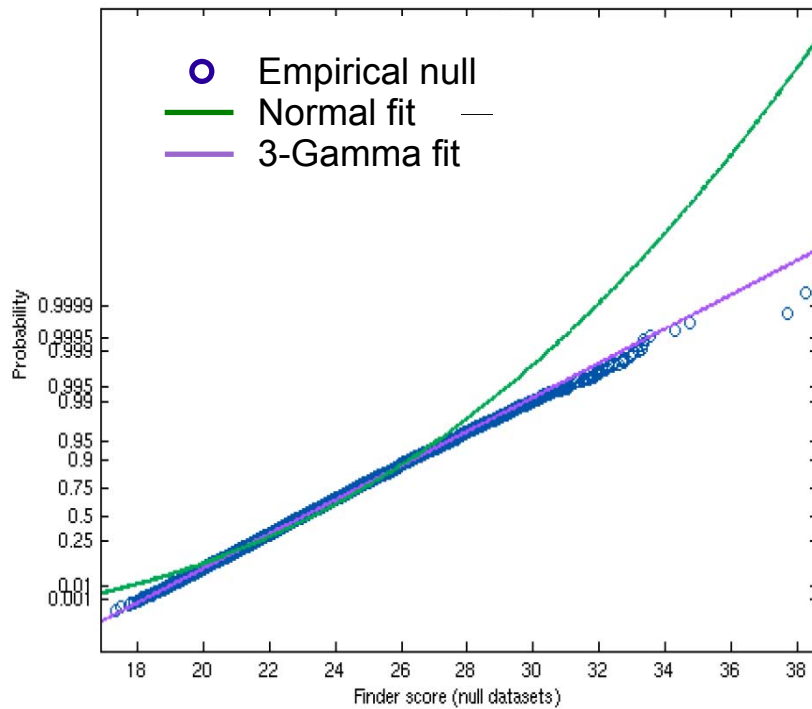
Is the Normal distribution a good approximation?



Empirical null is generated by applying GIMSA (Gibbs sampler) to 10,000 random sequence-sets from *S. cerevisiae* intergenic region. GIMSA evaluates motif based on log likelihood ratio (LLR) score.

6

3-Gamma vs Normal



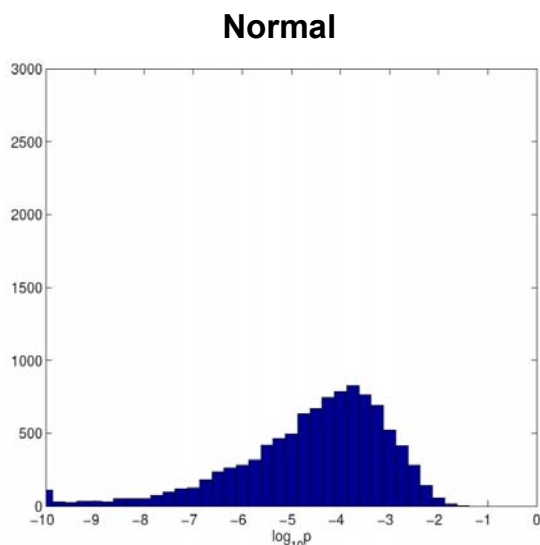
3-Gamma distribution is a good fit while the normal distribution is not

7

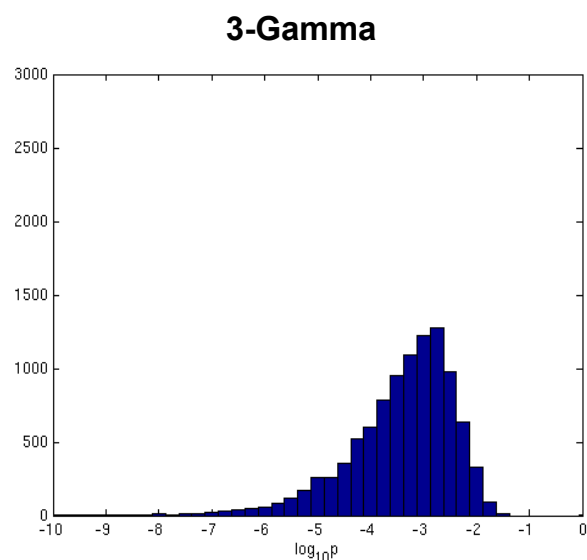
Normal approximation overestimates the significance

Histograms of 10,000 evaluations of the point estimator for p -value = 0.001.

The data was resampled from the HSA1 intergenic region.



\hat{p}_n overestimates the significance

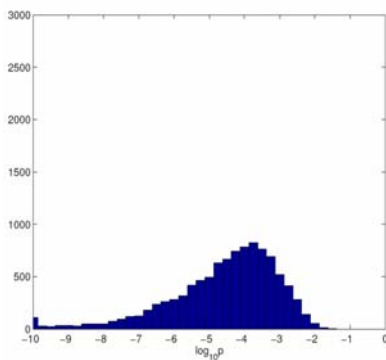


An ideal estimator $p(s)$ should have all the mass on the point $\log(p) = -3$

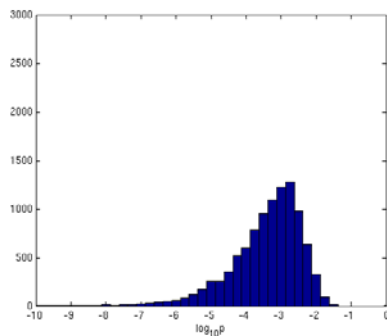
8

Can we do better?

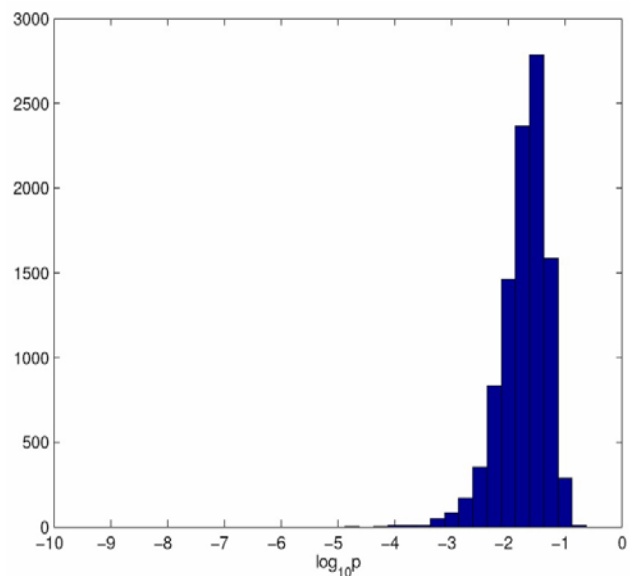
Normal p -value point estimator



3-Gamma p -value point estimator



3-Gamma “Confidence p -value”



(b) $\hat{p}_c(s, X)$ is mostly conservative

An ideal estimator $p(s)$ should have all the mass on the point $\log(p) = -3$

9

“Random” set of DNA sequences

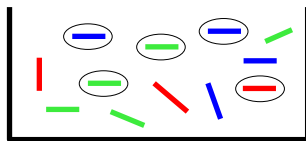
- Is the predicted motif from my input data special comparing to a motif I found in a **random** set of sequences?
- What is a “random” set of DNA sequences?
 - Uniformly sampled from a genome (widely used technique)
 - For example, from *S. cerevisiae* intergenic region
 - Harbison *et al.* 2004, Liu *et al.* 2001, Narlikar *et al.* 2007
 - High-order Markov Model
 - Shuffling

Local Sequence Composition

S. cerevisiae genome
> 13 million bp



Divide into overlapping
windows of 50bp



a sequence from input set
< 2000 bp

AATGGTTGGTCAAGGATGCGCAACCCAAATGATTCCTTGTTCCTCATTATTCTGGACA

Divide into non-overlapping
windows of 50bp



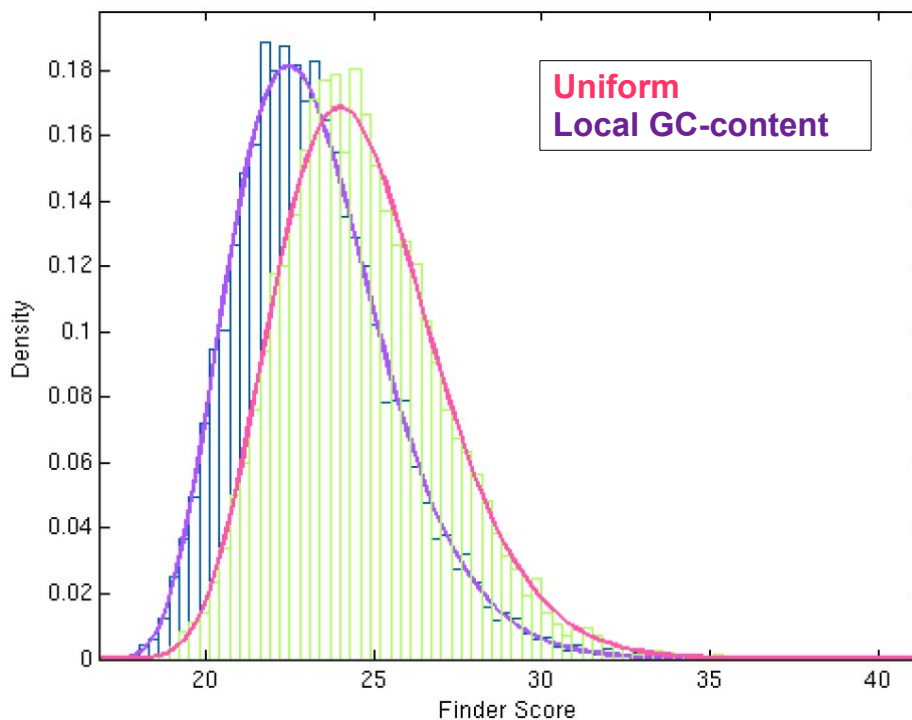
Replace each original window with a randomly drawn window
from a bin with similar local composition (e.g. GC-content)



GAGAACGCGTGGAGACAGATCAAATCTCAGCAGCAGATGTTGTTATGTTATC

11

Does preserving local GC-content make a difference?



Comparing empirical null distributions of log likelihood ratio (LLR) score from GIMSAN
Data from *S. cerevisiae* intergenic region

12

GIMSAN (Web Application)

- GIMSAN (GibbsMarkov with Significance ANalysis) is a novel web application
 - Gibbs sampler for motif finding
 - Biologically realistic and reliable statistical significance analysis

Job name: (please, no spaces, special characters etc., underscore is OK)

Upload your input FASTA file

Estimate background model from

your own genomic file (recommended)
 one of our standard genomic files
 input FASTA file

S288 S. cerevisiae entire genomic content
S288 S. cerevisiae intergenic content
H. Sapiens chr1 intergenic content

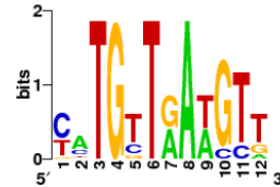
Width parameters (comma-separated list of integers)

Size of the null set

Program options

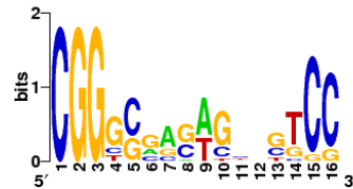
| | |
|--|--|
| Zero/one occurrences per sequence | |
| <input checked="" type="checkbox"/> -zoops | <input type="text" value="0.2"/> |
| Single-process time/cycles limit | Rapid convergence rate |
| -cput (seconds) <input type="text" value="300"/> | -L <input type="text" value="200"/> |
| Consider double strand | Order of Markov background |
| <input checked="" type="checkbox"/> -ds | -markov <input type="text" value="5"/> |

span: 12, logo constructed from 17 sequences
The MLE of the p-value is 0.12 and its confidence interval is (0, 0.44)



[Column pairs with statistically significant dependency \(0 pairs\)](#)
[Motif finder detailed output](#)

span: 16, logo constructed from 13 sequences
The MLE of the p-value is 0.0096 and its confidence interval is (0, 0.15)



[Column pairs with statistically significant dependency \(0 pairs\)](#)
[Motif finder detailed output](#)

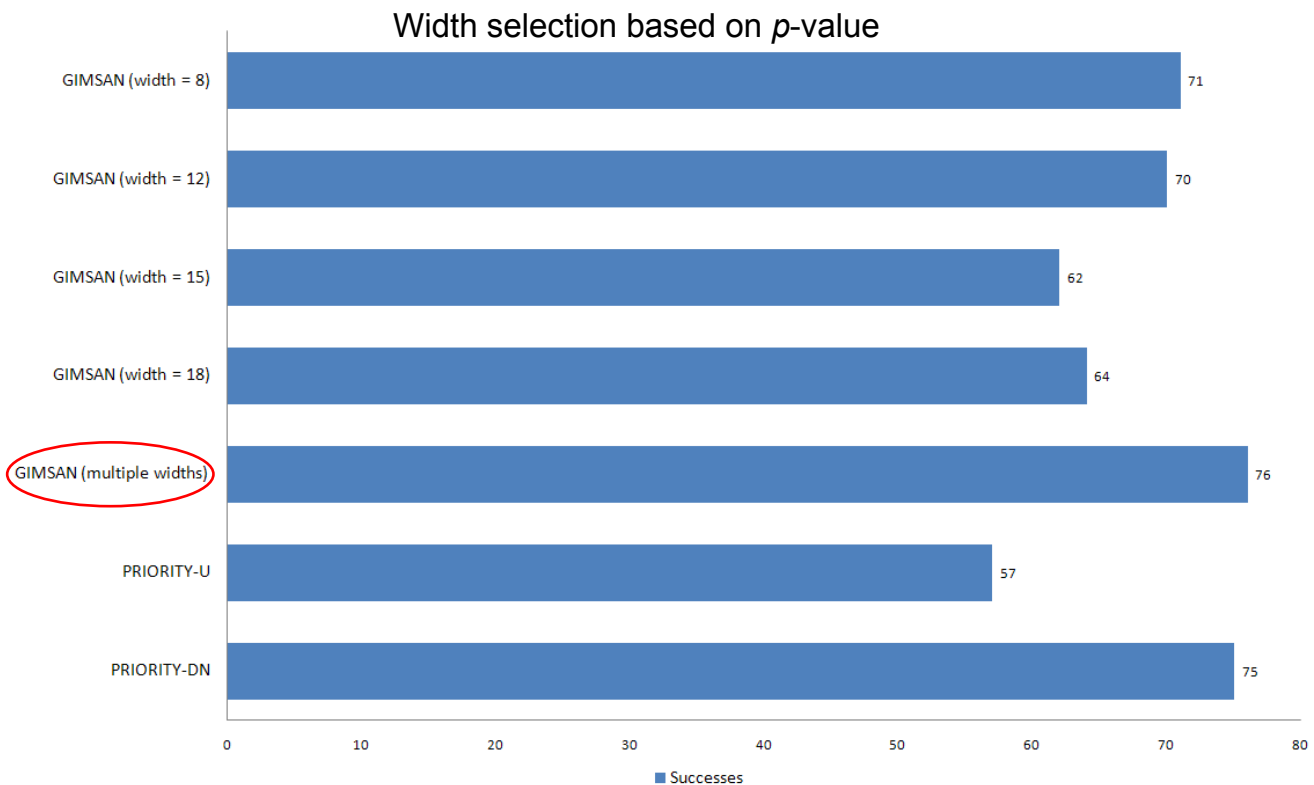
13

Results on Biological Data

- CHIP-chip data of transcription factors for *S. cerevisiae* from Harbison *et al.* 2004
 - Benchmark from Narlikar *et al.* 2007
 - 156 sequence-sets from 80 TFs
 - PRIORITY, AlignACE, MEME, MDscan, CONVERGE, conservation-based method of Kellis *et al.*
- How are successes defined?
 - Average mean-square-error

14

Using p -values to improve performance



Competitive with PRIORITY-DN (best finder from Narlikar 2007) 15

How well calibrated are our p -values?

- How well calibrated are our 3-Gamma confidence p -values on **biological data**?
- Perform hypothesis testing with a p -value threshold
- Is the observed rate $FP / (FP + TN)$ consistent with the theoretical rate?
 - CHIP-chip data of *S. cerevisiae* TFs from Narlikar *et al.*
 - At a 5% p -value threshold, observed rate of our p -value is 11% while the rate of its normal approximation counterpart is 74% (6-fold)
 - False-positive correction: motifs that have matches in database but does not match literature
 - At a 5% p -value threshold, observed rate is 8.3%
 - At a 10% p -value threshold, observed rate is 13.9%

Summary & Future Research

- 3-Gamma approximation
 - Much better fit than normal approximation for motif significance evaluation
- Local sequence composition
 - Framework for factoring biologically realistic information
 - Can be extended to factor other features
 - Use HMM model rather than our window-based method
- GIMSAN available as a web application
 - Competitive to other motif-finders that use additional information (e.g. conservation, nucleosome positioning)
 - Our p -values are well calibrated on biological data

17

Acknowledgment

- I would like to thank my advisor Uri Keich
- Anand Bhaskar for data processing
- NIH grant 1S10RR020889
- NSF grant No. 0644136

18