Gibbs Sampling for Gene Cis-Regulatory Elements

Lee A. Newberg @GE 10/21/2011
Acknowledgments

Team:
- Sean P. Conlan (NIH)
- Travis J. Desell (UND)
- Charles E. Lawrence (Brown)
- Lee Ann McCue (Pacific Northwest NL)
- Thomas M. Smith (MIT Lincoln Laboratory)
- William A. Thompson (Brown)

Resources:
- Wadsworth Center (including LCSB Core Facility)
- Rensselaer Polytechnic Institute
- Brown University (including CCMB)
- NIH/NHGRI: K25 Mentored Career Award “Quantitative Cross-Species Approaches to Gene Regulation” (LAN)
- DOE: “Bayesian computational approaches for gene regulation studies of bioethanol and biohydrogen production” (CEL, LAM, LAN)
Gibbs sampling for Gene Cis-Regulatory Elements
Gibbs sampling for gene regulation

- DNA *cis*-regulatory elements
- Importance
- Computational Prediction: Inputs, Outputs
- Results
- Algorithm
  - Statistical model, Gibbs sampling, & centroids
DNA cis-regulatory elements

Importance of DNA cis-regulatory elements

Important … for the understanding of cell function, differentiation, and pathology

… because the elements affect both the products of genes and when and to what extent the genes are expressed

Typically vary species to species, but not individual to individual, except pathologically.

Howard-Ashby, Materna, Brown, Tu, Oliveri, Cameron, & Davidson, *Dev Biol*, 2006

Lee A. Newberg
10/21/2011

Gibbs sampling for Gene Cis-Regulatory Elements
Outputs

- Element sites
- Motif (pattern) description

Input data

- Promoter sequences: aligned when feasible (e.g., 20 sequences × 5 species)
- Phylogenetic tree and model: or ad hoc substitute

Input parameters

- Motif model type: consensus w/ deviations vs. probabilistic
- Motif size: fixed, varying. (6-36 nts.)
- Motif shapes: palindromic, off positions
- Site frequency: per promoter, genome
- Site positioning: nucleosomes, relative to +1
Effectiveness in simulations

Verifying known results

Novel predictions (by others)

E.g.: Driscoll et al. (2007) Carbon utilization pathway in Shewanella
What we do, part 1 of 3: Statistical Modeling

- **Plausible, workable**, statistical model components:
  - Position weight matrices for motif models, etc.

\[
Q = \left( \begin{array}{ccccc}
* & \pi_C & \kappa \pi_G & \pi_T \\
\pi_A & * & \pi_G & \kappa \pi_T \\
\kappa \pi_A & \pi_C & * & \pi_T \\
\pi_A & \kappa \pi_C & \pi_G & * \\
\end{array} \right)
\]
What we do, part 2 of 3: Gibbs sampling (MCMC)

Random walk through posterior probability space. Possible because: we resample part of a solution, conditioned on the rest.

- Element sites
- Motif model logos
- Tree sequence alignment, etc.

Record key features as we walk
What we do, part 3 of 3: Centroid solution

We focus on the region of solution space containing the most posterior probability, rather than on the single solution that is most probable.

→ Build centroid from marginal probability of relevant features
C/C++ UNIX.

Single CPU / Small Cluster / DNA@home.
Payoff

Plausible statistical model components
+ Gibbs sampling
+ Centroid
→ Robust predictions

- Theory: Newberg et al. (2007), Bioinformatics, Pubmed17488758
- Use: Thompson et al. (2007), Nucleic Acids Research, Pubmed17483517