DEVELOPING STATISTICAL AND ALGORITHMIC METHODS FOR SHOTGUN METAGENOMICS AND TIME SERIES ANALYSIS

– review of my PhD work

Li Charlie Xia

Molecular and Computational Biology, University of Southern California

April 23, 2012

Outline

- Introduction
  - Biological Background
  - Our Computational Approaches
- GRAMMy for Shotgun Metagenomics
  - Motivations
  - Methods
  - Simulation Results
  - Application to Real Datasets
  - Conclusions
- eLSA for Time Series Data
  - Biological Background
  - Methods
  - Simulation Results
  - Application to Real Datasets
  - Conclusions
- Future work
- References

Outline

- Introduction
  - Biological Background
  - Our Computational Approaches
- GRAMMy for Shotgun Metagenomics
  - Motivations
  - Methods
  - Simulation Results
  - Application to Real Datasets
  - Conclusions
- eLSA for Time Series Data
  - Biological Background
  - Methods
  - Simulation Results
  - Application to Real Datasets
  - Conclusions
- Future work
- References

High-throughput (HT) Molecular Technologies

High-throughput Molecular Technologies

Researchers can study hundreds of bio-molecules simultaneously.

- DNA: Massive shotgun sequencing *Venter et al. (2004)*
- Proteins: Protein microarray *Zhu et al. (2001)*
- RNAs: Microarray, RNA-Seq *Augenlicht and Kobrin (1982); Wang et al. (2009)*

'omics

These HT technologies make possible many 'omics study.

- Biological Pathway Systems: Proteomics
- Transcriptional Regulation: Transcriptomics
- Microbial Communities: Metagenomics

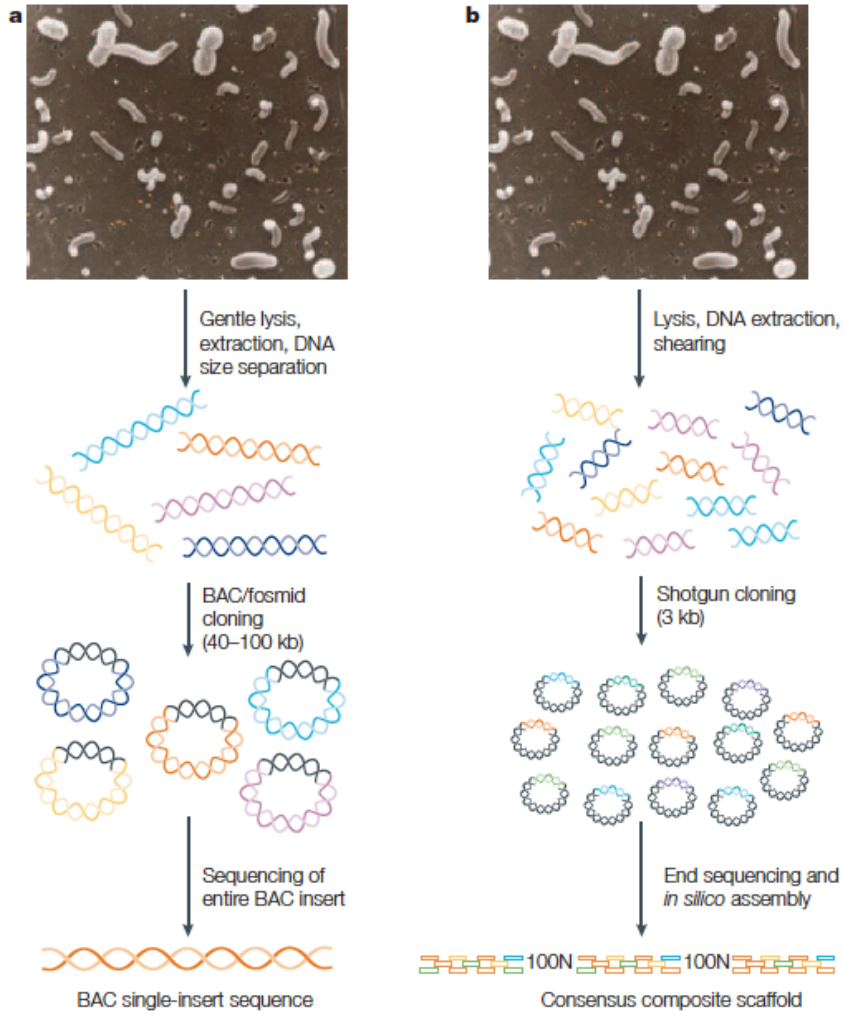
Shotgun Metagenomics

Metagenomics

The application of modern genomics techniques to the study of communities of microbial organisms directly in their natural environments, bypassing the need for isolation and lab cultivation of individual species. *Handelsman et al. (1998)*.

Shotgun Metagenomics

Researchers sample uncultured microorganisms, randomly shear DNA, sequence many short reads. Figure from *DeLong (2005)*.



Our Computational Approaches

To meet the challenge for computational analysis of new HT data.

GRAMMy for shotgun metagenomics Xia et al. (2011a)

- Accurate and efficient estimation of relative abundance.

eLSA for time series data Xia et al. (2011b)

- Capture local and time-delayed associations

Outline

- Introduction
  - Biological Background
  - Our Computational Approaches
- GRAMMy for Shotgun Metagenomics
  - Motivations
  - Methods
  - Simulation Results
  - Application to Real Datasets
  - Conclusions
- eLSA for Time Series Data
  - Biological Background
  - Methods
  - Simulation Results
  - Application to Real Datasets
  - Conclusions
- Future work
- References

Motivations

Biological Question

To estimate the microbial community composition more accurately based on the shotgun metagenomic reads.

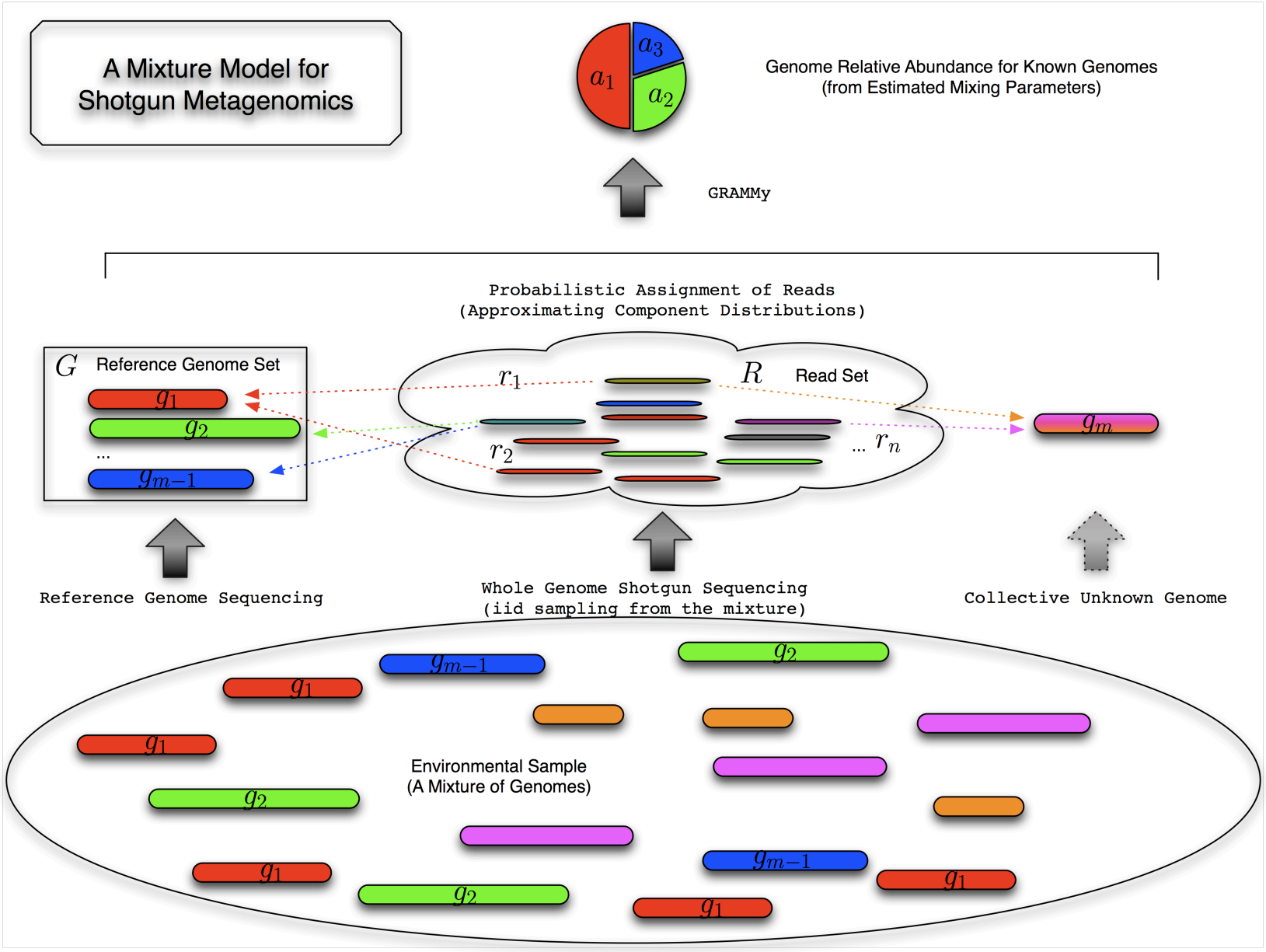
Motivations

- Increasingly available short read datasets.
- Increased read assignment ambiguities.
- Rich abundance information in NGS read sets.
- Future metagenomics:
  (1) NGS read set: high coverage.
  (2) Reference set: database and single cell sequencing.

The GRAMMy Framework

GRAMMy: Genome Relative Abundance(GRA) estimation using Mixture Model theory(MMy).

The GRAMMy Model



Model Definitions

The Genome Mixture:

$$\mathcal{M} : \mathcal{M} = \sum_{j=1}^{m} \pi_j f_{g_j}$$

Genome Relative Abundance (GRA):

$$a_j = \frac{\#\text{ of } j\text{-th known unique genome}}{\#\text{ of all known genomes}},$$

where $\sum_{j=1}^{m-1} a_j = 1$.

Model Definitions

Relating Mixing Parameters to GRA:

$$a_j = \frac{\pi_j}{l_j \sum_{k=1}^{m-1} \frac{\pi_k}{l_k}}, \tag{1}$$

and the inverse form:

$$\pi_j = c \cdot \frac{a_j l_j}{\sum_{k=1}^{m-1} a_k l_k}, \tag{2}$$

where $j = 1, 2, ..., m-1$ and $c = 1 - \pi_m$.

The GRAMMy EM Algorithm

E-step:

$$z_{ij}^{(t)} = \frac{p(r_i|z_{ij}=1;\mathbf{G})\pi_j^{(t)}}{\sum_{k=1}^{m} p(r_i|z_{ik}=1;\mathbf{G})\pi_k^{(t)}}. \tag{3}$$

M-step:

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^{n} z_{ij}^{(t)}}{n}. \tag{4}$$

Approximate Component Distribution

f_{g_j} using mapping ('map')

The number of high quality mapping or alignment hits s_{ij} for read r_i on genome g_j can be used:

$$p(r_i|z_{ij} = 1; \mathbf{G}) \propto \frac{s_{ij}}{l_j}. \tag{5}$$

f_{g_j} using k-mer ('k-mer')

k-mer composition based on multinomial distributions can be used:

$$p(r_i|z_{ij} = 1; \mathbf{G}) \propto \prod_{w \in W_i} p_{wj}. \tag{6}$$

$W_i = \{\text{sliding window k-mers of } r_i\}$ and $p_{wj} = \frac{\#\text{ of } w \text{ in } g_j}{l_j}$.

Standard Error Estimation

Standard Error from Asymptotic Theory

When the mapped read number is large, we can use:

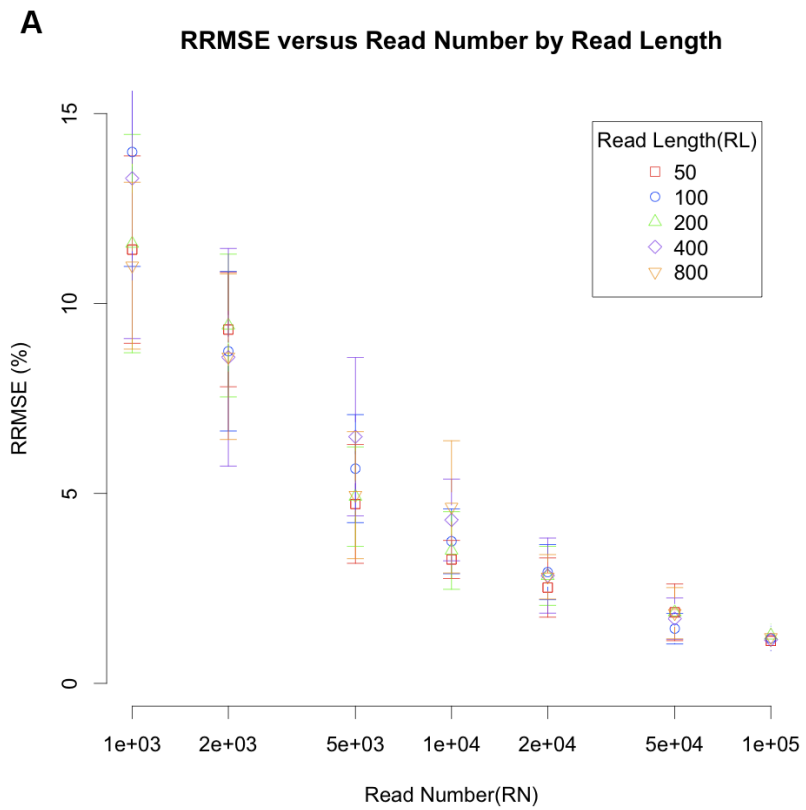
$$SE(a_j^*) = (\mathbf{Cov}(\mathbf{a}^*))_{jj} \approx ((\mathbf{I_o^{-1}}(\mathbf{a}|\mathbf{R},\mathbf{G}))_{jj})^{\frac{1}{2}}\Big|_{\hat{\pi}=\hat{\pi}^*}. \quad (7)$$

Standard Error from Empirical Distribution

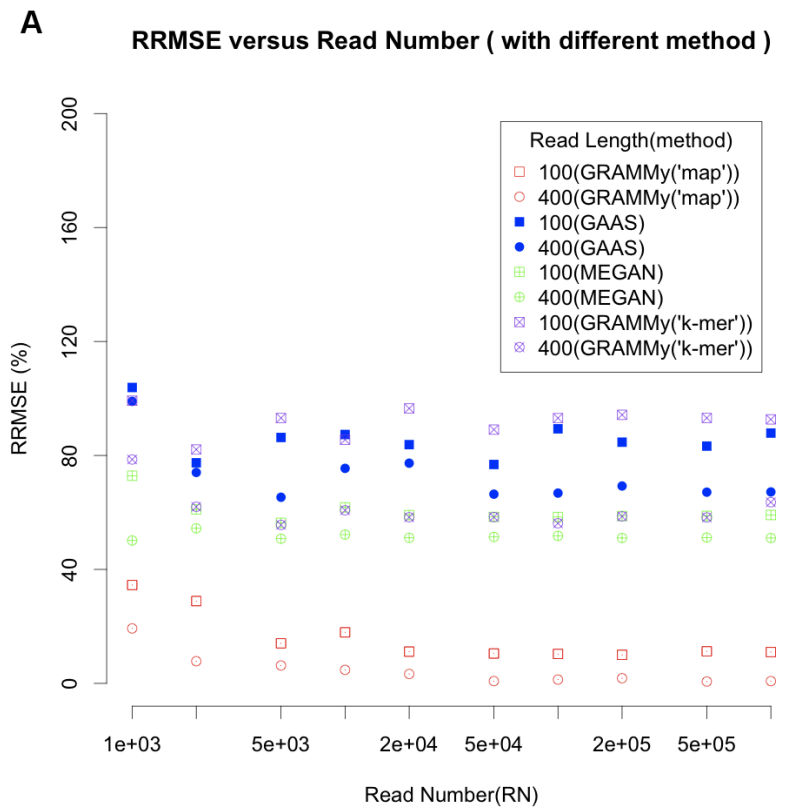
When the mapped read number is small, we can use:

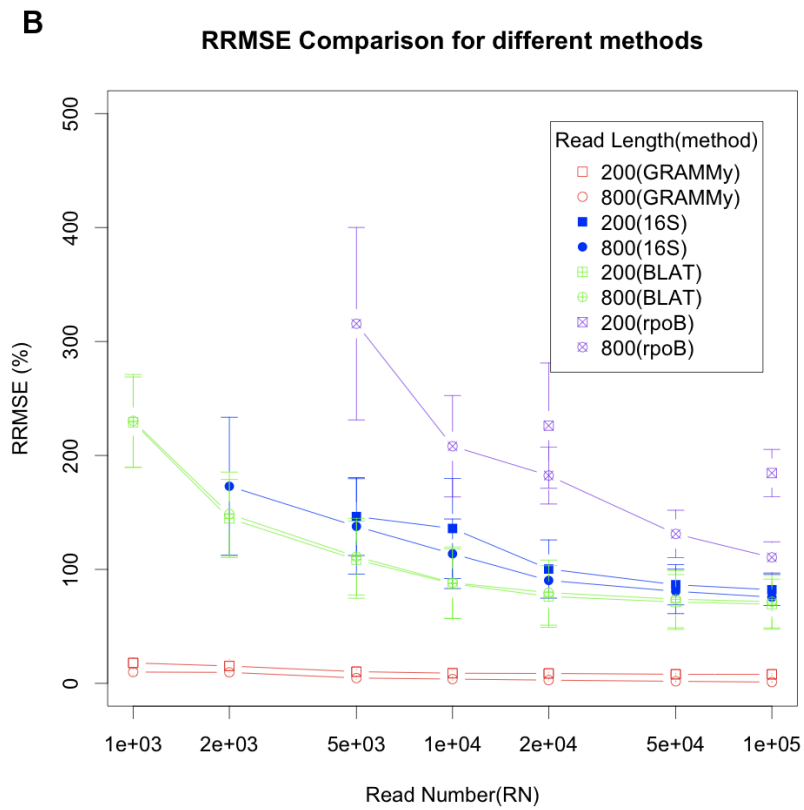
$$SE(a_j^*) = (\mathbf{Cov}(\mathbf{a}^*))_{jj} \approx (\frac{1}{B-1}\sum_{b=1}^{B}(\mathbf{a}_{(b)}^* - \bar{\mathbf{a}^*})(\mathbf{a}_{(b)}^* - \bar{\mathbf{a}^*})^T)_{jj}^{\frac{1}{2}}. \quad (8)$$

Simulation: The Convergence of EM Estimates

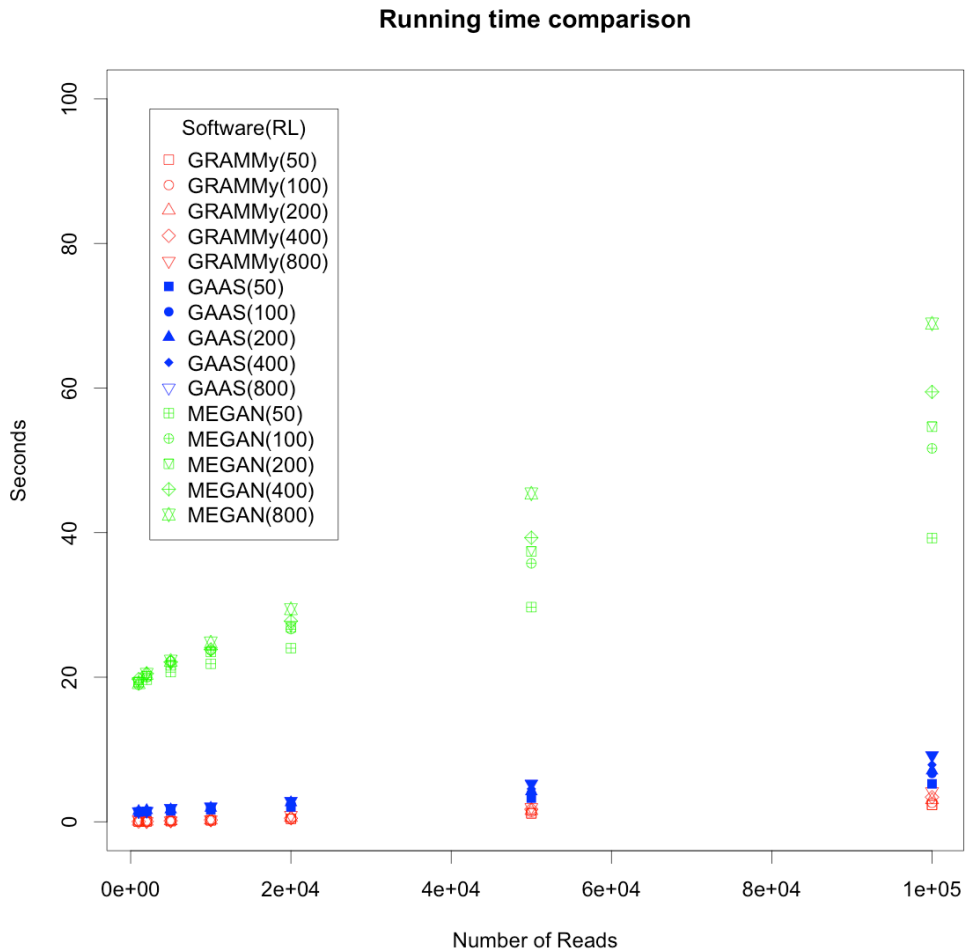


Simulation: Comparison with other Methods





Running time comparison



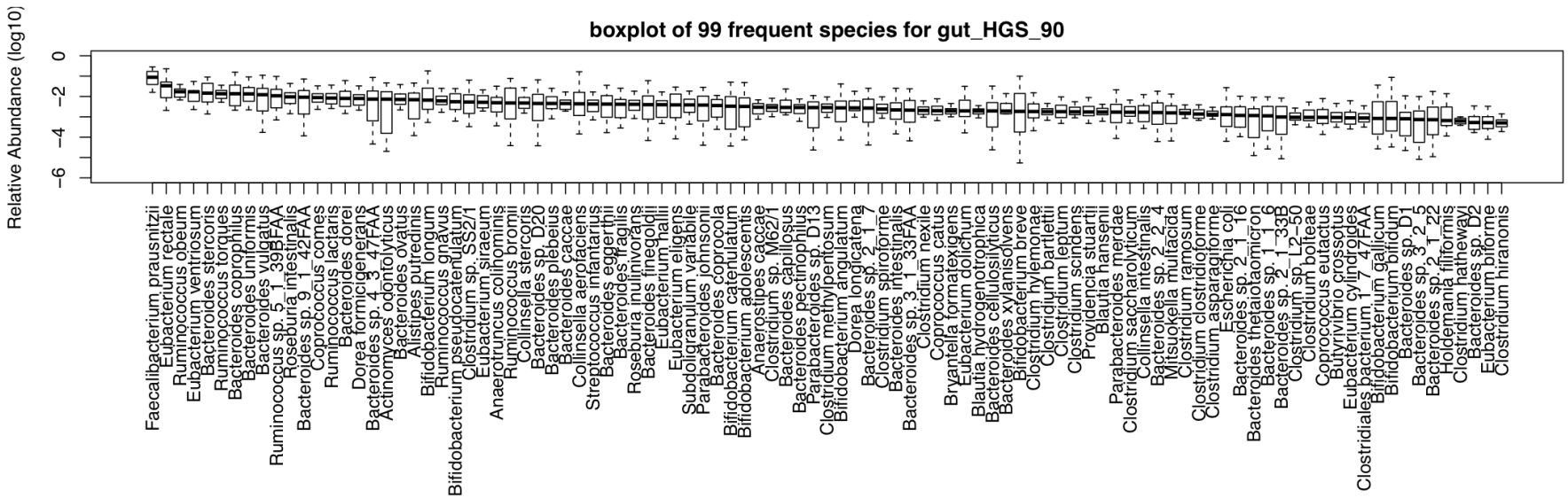
Artificial metagenomes with real reads

	simLC		simMC		simHC	
	RRMSE	AVGRE	RRMSE	AVGRE	RRMSE	AVGRE
GRAMMy	20.0%	14.0%	25.6%	19.7%	21.6%	14.7%
MEGAN	48.6%	39.3%	50.0%	40.6%	50.2%	40.8%
GAAS	433.8%	152.5%	171.4%	111.6%	507.9%	165.8%

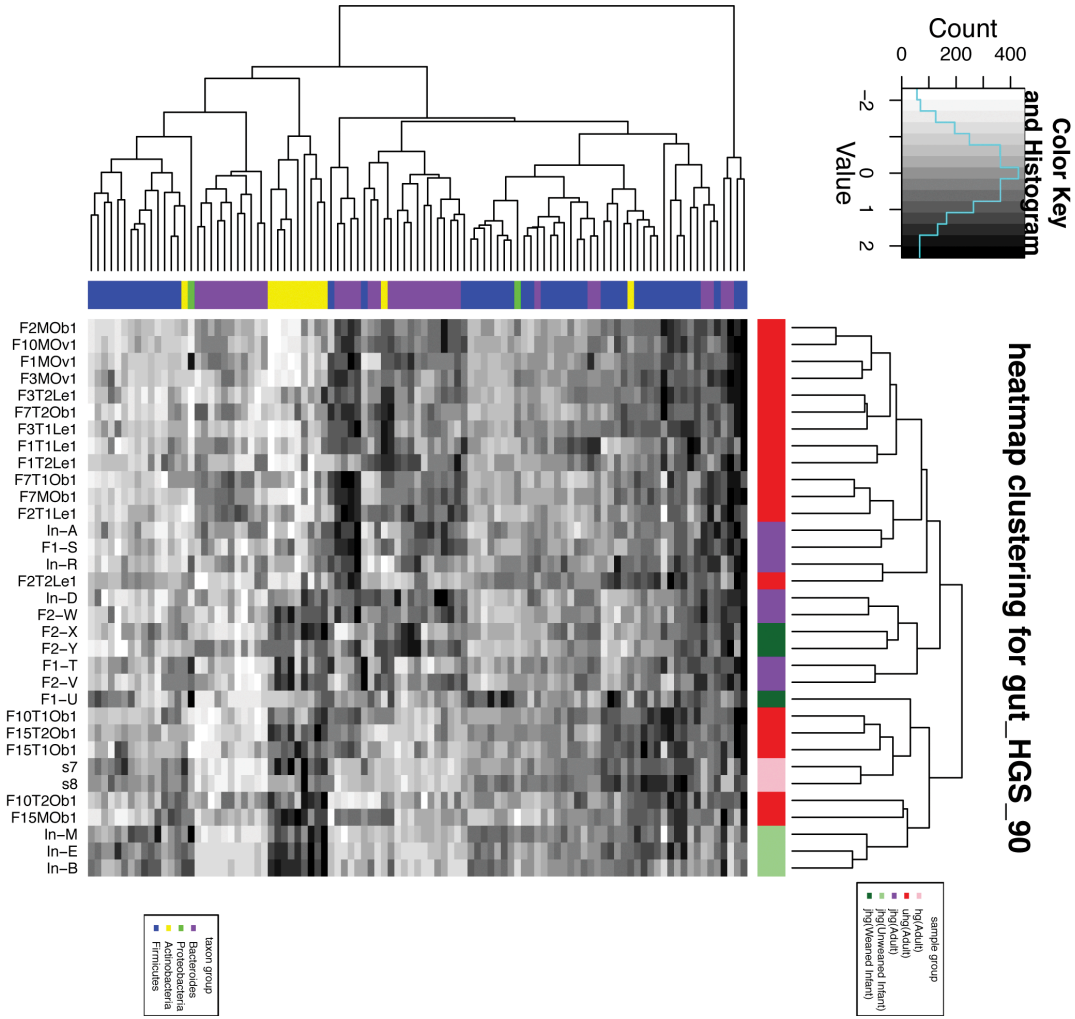
Summary Statistics for Real Datasets

	Mapped Rate(%)			Ambiguity Rate(%)			Average Genome Length(bp)		
Data Set(#Sets)	Med.	Min.	Max.	Med.	Min.	Max.	Med.	Min.	Max.
hg_HGS(2)	46.65	43.15	50.15	31.65	30.32	32.98	2890092	2660792	3119393
jhg_HGS(13)	59.61	35.99	76.92	45.11	22.53	65.71	3745629	2268438	5657331
uhg_HGS(18)	52.35	37.49	72.51	35.90	21.56	59.81	3619072	3047940	4752910
amd_AMD(1)	45.64	45.64	45.64	1.48	1.48	1.48	2163584	2163584	2163584

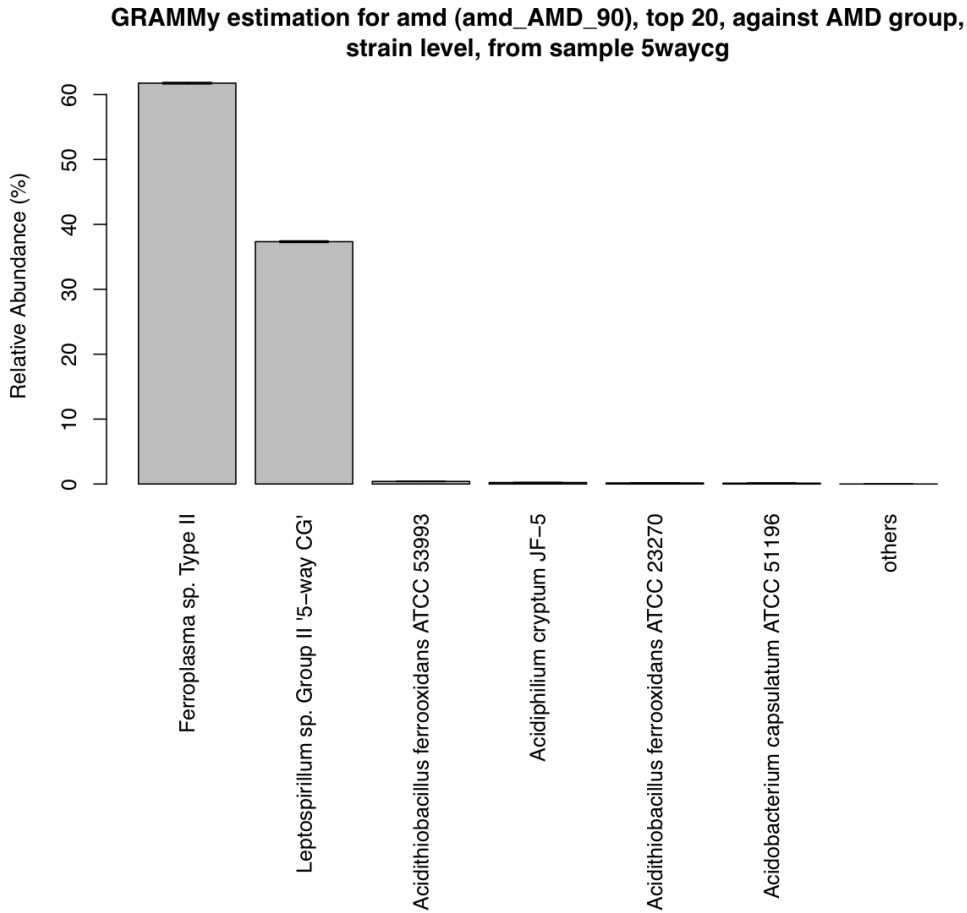
Human Gut Frequent Species



Human Gut Bi-clustering



Acid Mine Drainage Community



Conclusions

Features of GRAMMy

- Explicitly model read assignment probabilities.
- Can make use of references as well as other assemblages.
- Choices of read assignment methods.
- More accurate and faster compared to existing methods.

Availability

The GRAMMy software is available from GRAMMy's homepage at `http://meta.usc.edu/softs/grammy`

Outline

- Introduction
  - Biological Background
  - Our Computational Approaches
- GRAMMy for Shotgun Metagenomics
  - Motivations
  - Methods
  - Simulation Results
  - Application to Real Datasets
  - Conclusions
- eLSA for Time Series Data
  - Biological Background
  - Methods
  - Simulation Results
  - Application to Real Datasets
  - Conclusions
- Future work
- References

Biological time series data

Biological Time Series

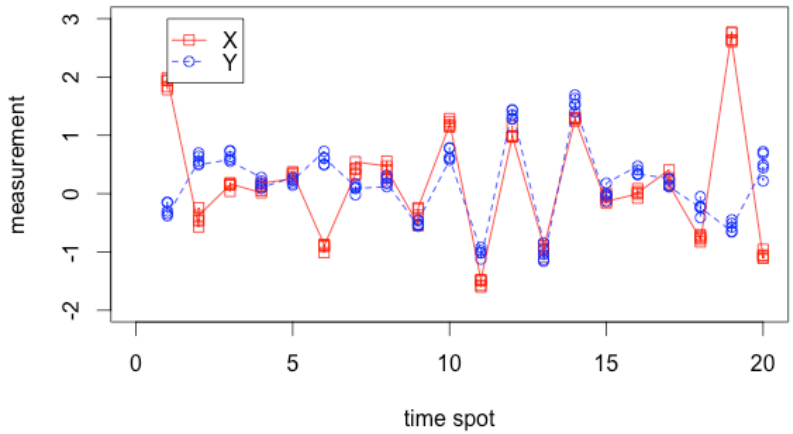
Transcriptomics study

- Microarray time series
  Gene expression study. *Wang and Kim (2003)*
- RNA-Seq time series
  Alternative splicing study. *Trapnell et al. (2010)*

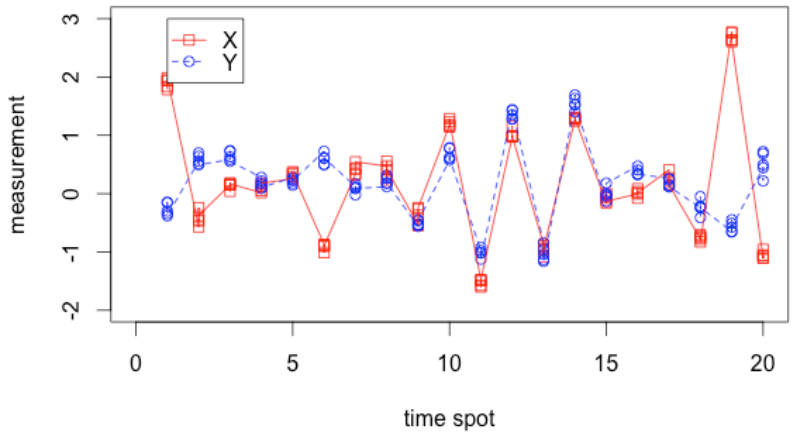
Biological Questions

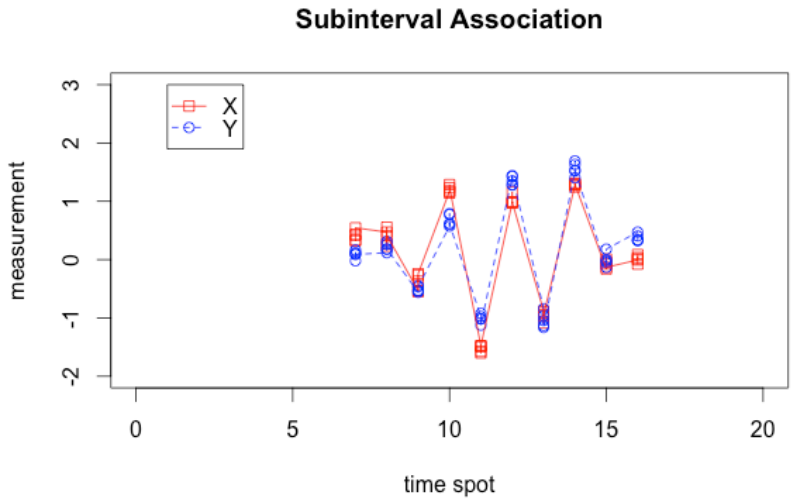
Natural assumption: associated genes and transcripts are likely coordinated to perform common functions.
It is important to identify which genes/transcripts (factors) are associated and how ?

Genes X and Y are associated or not?

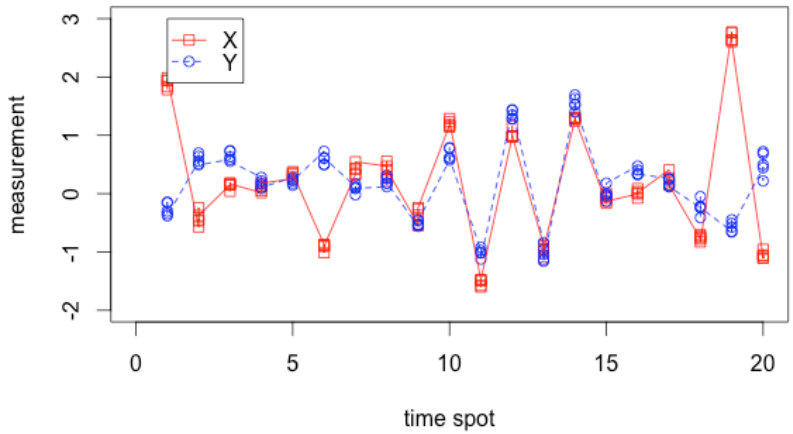


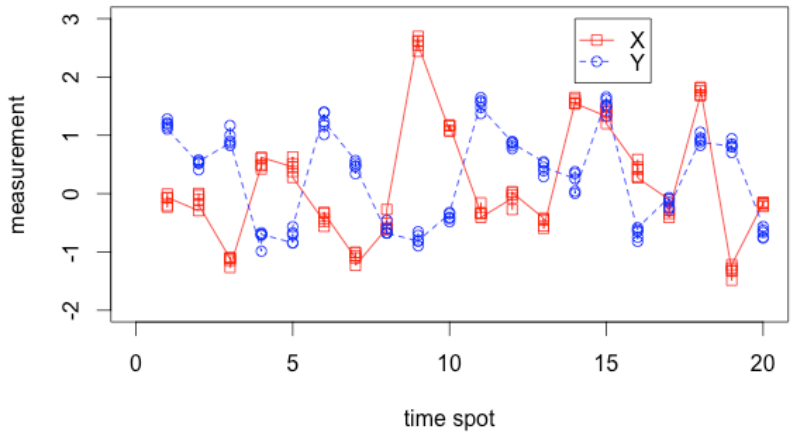
Genes X and Y are associated or not?

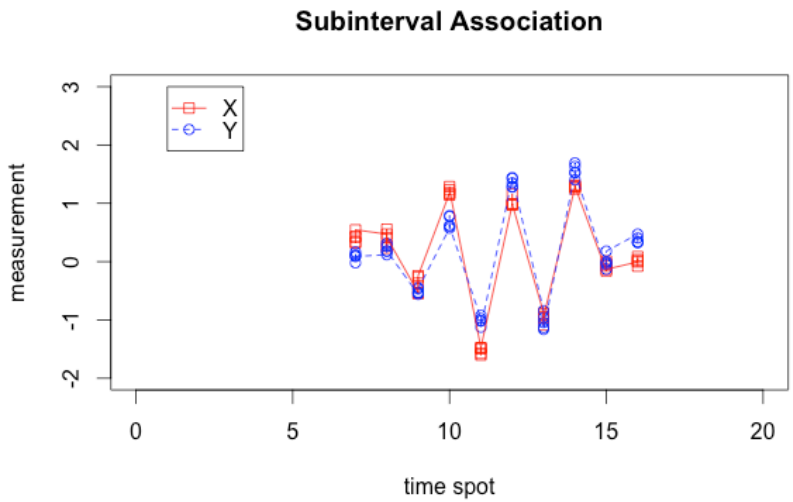




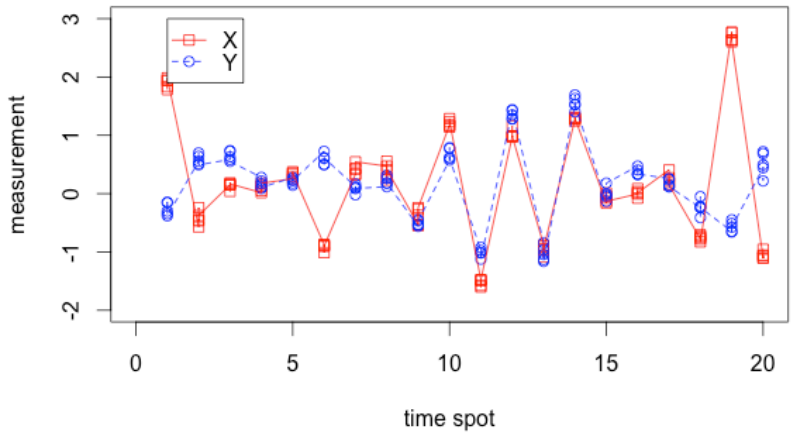
Genes X and Y are associated or not?

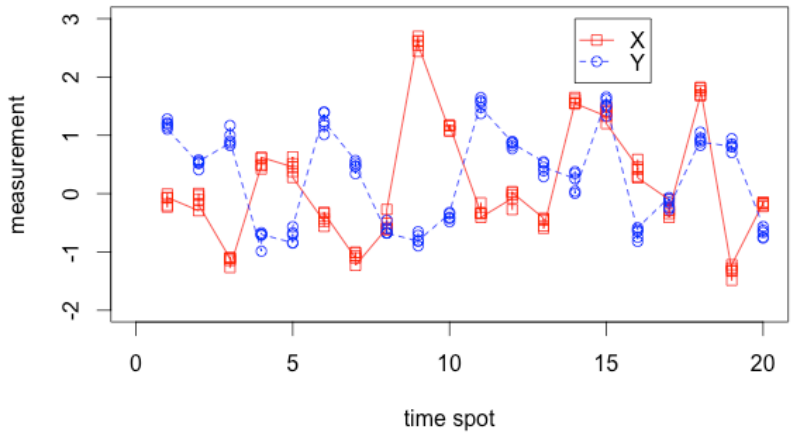


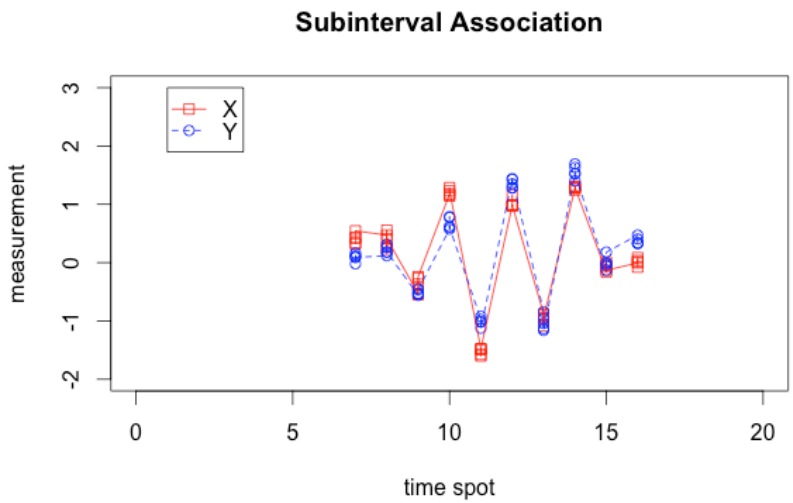


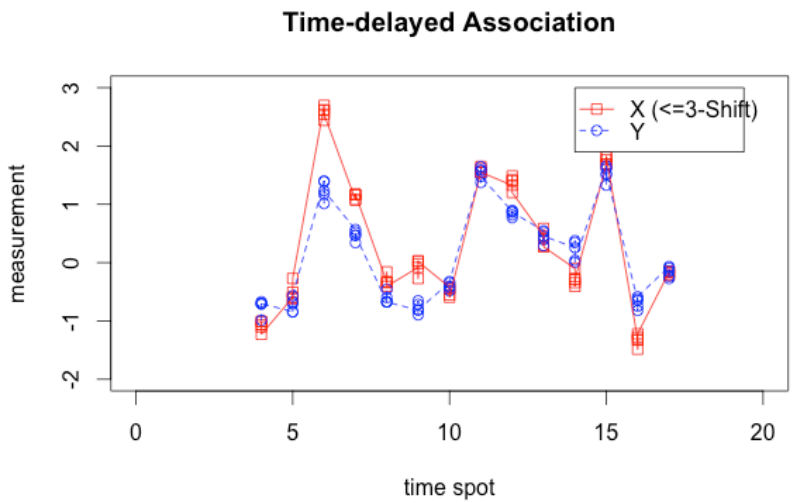


Genes X and Y are associated or not?









Local Similarity Analysis

Conventional Methods

Pearson's Correlation Coefficient (PCC) and others

- Limitations: difficulty in complex associations

Local Similarity Analysis (LSA)

Originally in *Ruan et al. (2006a)* . Ideas from local alignment – *Waterman (1995);* **?** . Find possibly shifted interval maximize the Local Similarity metric.

Local Similarity Metric

$$|S_{max}(x_{[1:n]}, y_{[1:n]})| = \max_{i,j,l} |\sum_{k=0}^{l-1} x_{i+k} y_{j+k}|$$

- LS score: S_{max}/n
- Delay limit: $|i - j| \leq D$

Motivations for eLSA

Motivations

- Original LSA proved useful

  e.g. Paver and Kent (2010) and Shade et al. (2010)
- Temporal and spatial series data with replicates

  e.g. Dr. Cardon's lab
- Confidence Interval (CI) for LS score can be obtained
- Accelerate and pipeline the tools for larger scale analysis

Extended Local Similarity Analysis (eLSA)

Extends LSA to replicated time series – efficient and pipelined.

Modeling

Time series data

Pairs of factors: $X = X_{[1:m][1:n]}$ and $Y = Y_{[1:m][1:n]}$
n: columns = time points; m: rows = replicates at each time point

Extended LS Metric

$$|S_{max}(X_{[1:m][1:n]}, Y_{[1:m][1:n]})| = \max_{i,j,l} |\sum_{k=0}^{l-1} F(X_{[1:m],i+k}) F(Y_{[1:m],j+k})|$$

- LS score: S_{max}/n; Delay limit: $|i - j| \leq D$
- F: summarizing function for repeated measures.

Dynamic Programming

Find possibly shifted interval maximizing the extended LS metric

F – Summarizing Repeated Measures

F functions

Scalars x_i's in LSA *vs.* column vector functions $F(X_i)$ in eLSA.

- simple average ('simple'): $F(X_i) = \bar{X}_i$
- Standard Deviation weighted ('SD'):

$$F(X_i) = \frac{\bar{X}_i}{\sigma_{X_i}}$$

- Median ('Med'): $F(X_i) = Median(X_i)$
- Median Absolute Deviation weighted ('MAD'):

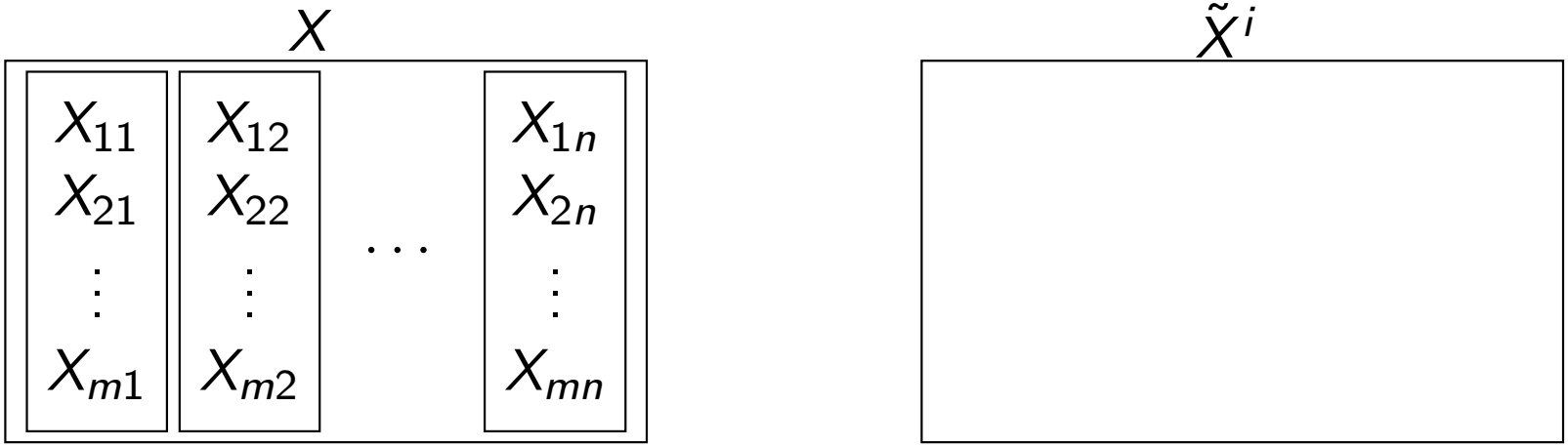
$$F(X_i) = \frac{Median(X_i)}{MAD(X_i)},$$

where $MAD(X_i) = Median(|X_i - Median(X_i)|)$

$F(X_i)$ – F-transformed data

Bootstrap CI for extended LS score

i-th Bootstrap sample $\tilde{X}^i$ (similarly $\tilde{Y}^i$)



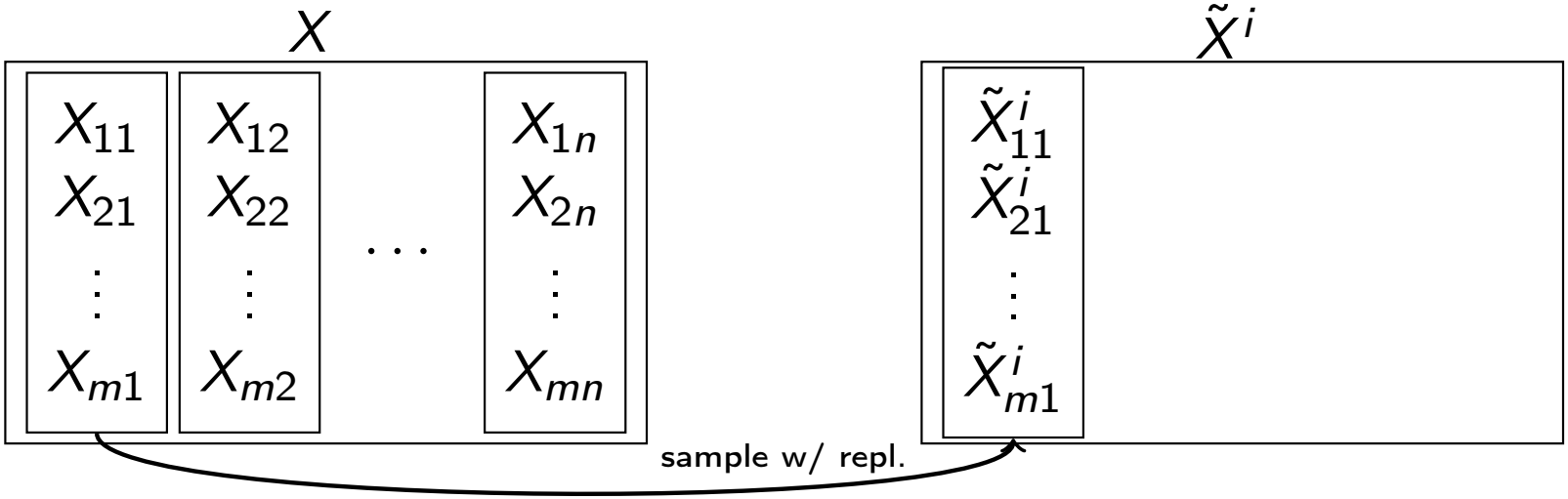
$1-\alpha$ Bootstrap CI

$\tilde{S}^{(j)}_{max}$: j-th largest of $S_{max}(\tilde{X}^i, \tilde{Y}^i)$'s; s.t. $\tilde{S}^{(1)}_{max} \leq \tilde{S}^{(2)}_{max} \leq \ldots \leq \tilde{S}^{(B)}_{max}$

$$\mathrm{CI}_{1-\alpha} = [\tilde{S}^{(\lfloor \frac{\alpha}{2} B \rfloor)}_{max}, \tilde{S}^{(\lfloor (1-\frac{\alpha}{2}) B \rfloor)}_{max}]$$

Bootstrap CI for extended LS score

i-th Bootstrap sample $\tilde{X}^i$ (similarly $\tilde{Y}^i$)



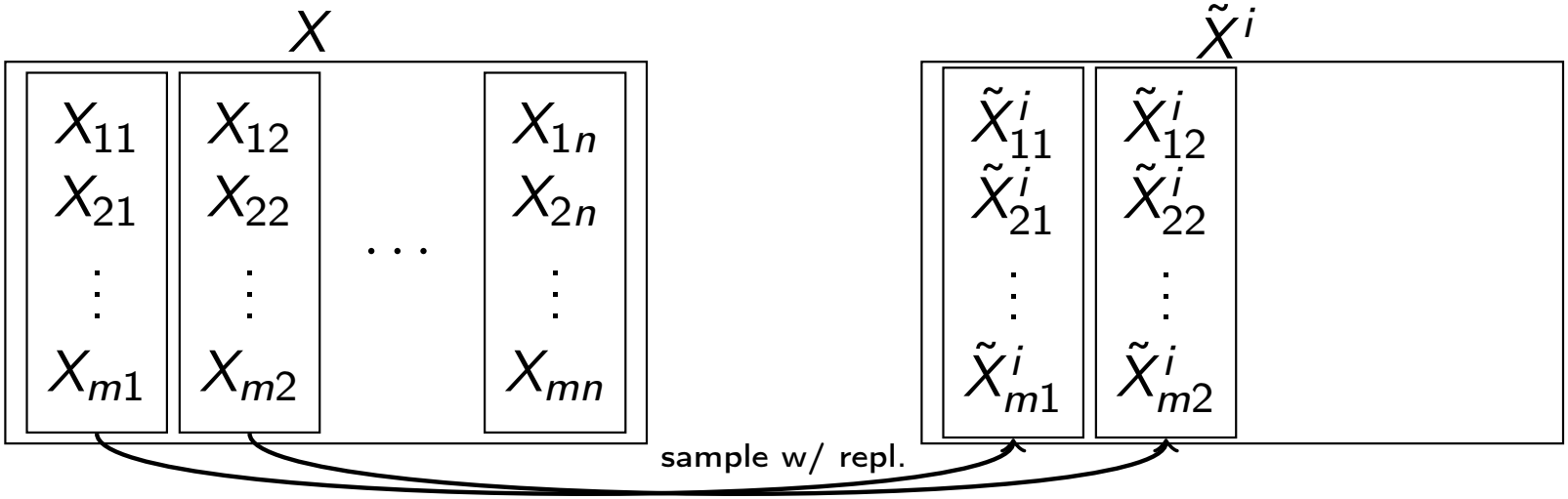
$1-\alpha$ Bootstrap CI

$\tilde{S}^{(j)}_{max}$: j-th largest of $S_{max}(\tilde{X}^i, \tilde{Y}^i)$'s; s.t. $\tilde{S}^{(1)}_{max} \leq \tilde{S}^{(2)}_{max} \leq \ldots \leq \tilde{S}^{(B)}_{max}$

$$\mathrm{CI}_{1-\alpha} = [\tilde{S}^{(\lfloor \frac{\alpha}{2} B \rfloor)}_{max}, \tilde{S}^{(\lfloor (1-\frac{\alpha}{2}) B \rfloor)}_{max}]$$

Bootstrap CI for extended LS score

i-th Bootstrap sample $\tilde{X}^i$ (similarly $\tilde{Y}^i$)



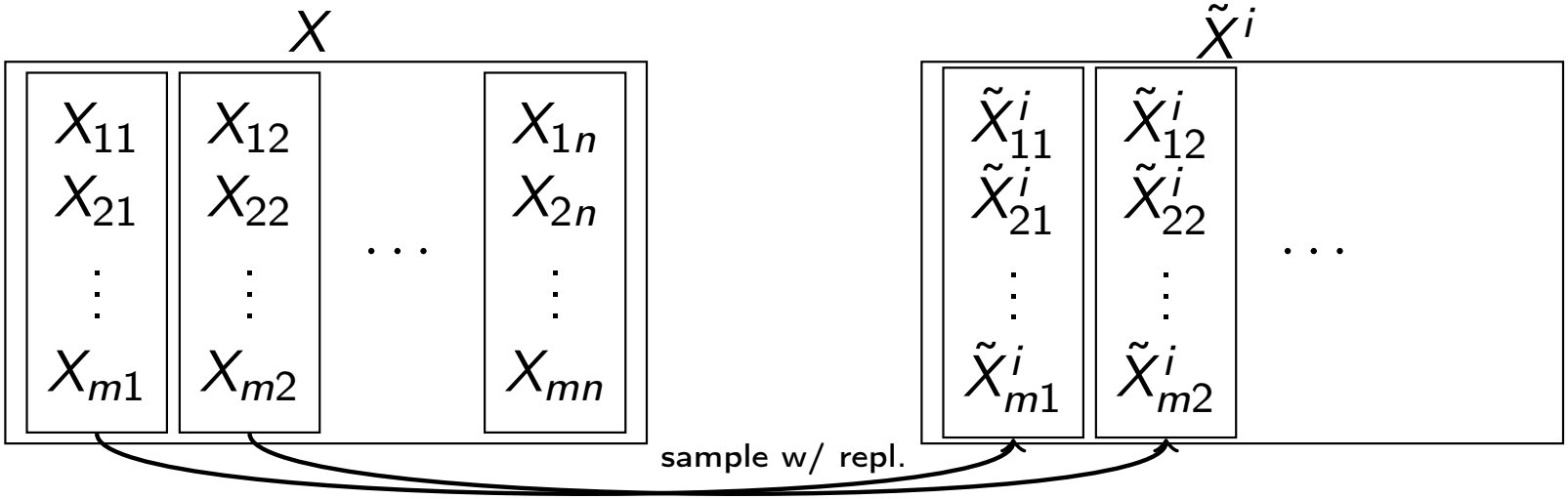
$1-\alpha$ Bootstrap CI

$\tilde{S}^{(j)}_{max}$: j-th largest of $S_{max}(\tilde{X}^i, \tilde{Y}^i)$'s; s.t. $\tilde{S}^{(1)}_{max} \leq \tilde{S}^{(2)}_{max} \leq ... \leq \tilde{S}^{(B)}_{max}$

$$\mathrm{CI}_{1-\alpha} = [\tilde{S}^{(\lfloor \frac{\alpha}{2} B \rfloor)}_{max}, \tilde{S}^{(\lfloor (1-\frac{\alpha}{2}) B \rfloor)}_{max}]$$

Bootstrap CI for extended LS score

i-th Bootstrap sample $\tilde{X}^i$ (similarly $\tilde{Y}^i$)



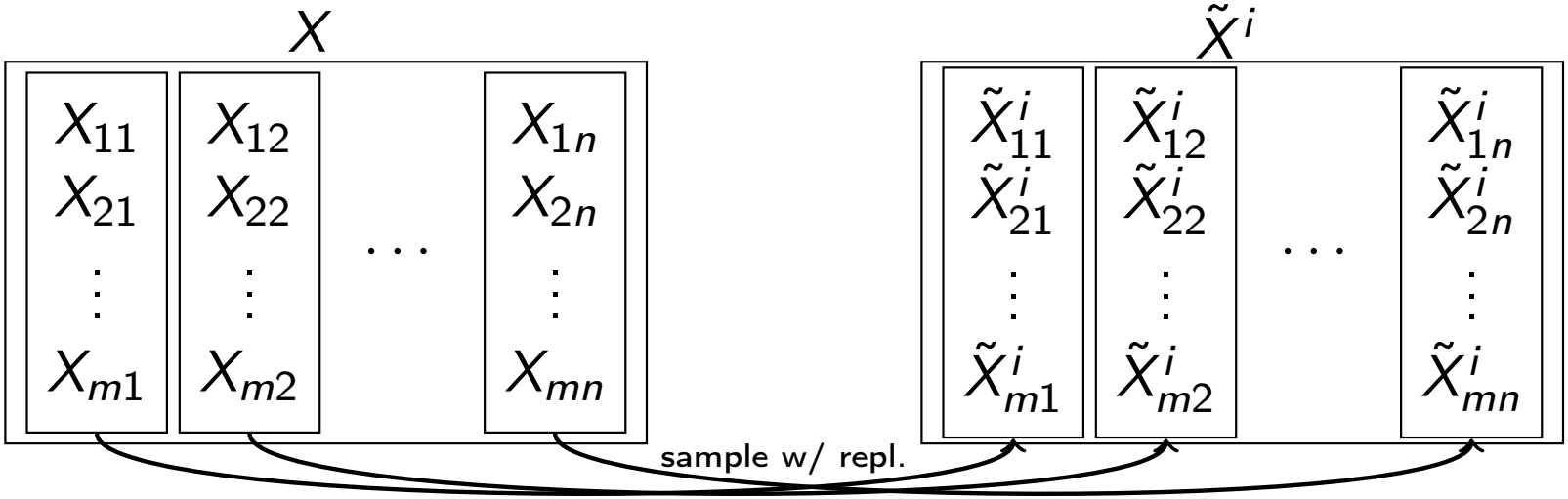
$1-\alpha$ Bootstrap CI

$\tilde{S}^{(j)}_{max}$: j-th largest of $S_{max}(\tilde{X}^i, \tilde{Y}^i)$'s; s.t. $\tilde{S}^{(1)}_{max} \leq \tilde{S}^{(2)}_{max} \leq ... \leq \tilde{S}^{(B)}_{max}$

$$\mathrm{CI}_{1-\alpha} = [\tilde{S}^{(\lfloor \frac{\alpha}{2} B \rfloor)}_{max}, \tilde{S}^{(\lfloor (1-\frac{\alpha}{2}) B \rfloor)}_{max}]$$

Bootstrap CI for extended LS score

i-th Bootstrap sample $\tilde{X}^i$ (similarly $\tilde{Y}^i$)



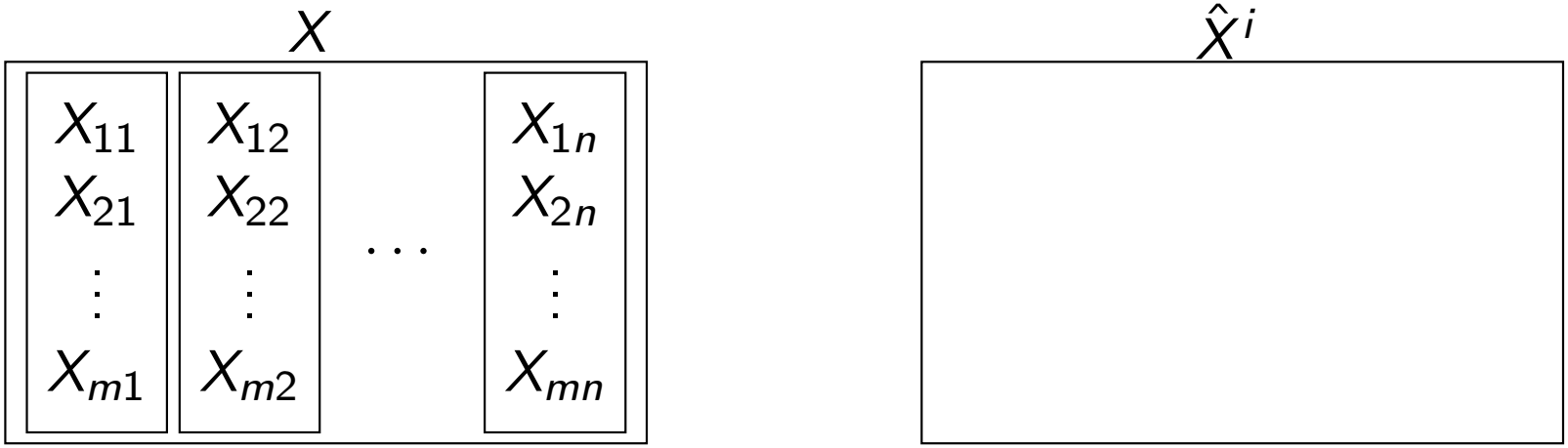
$1-\alpha$ Bootstrap CI

$\tilde{S}^{(j)}_{max}$: j-th largest of $S_{max}(\tilde{X}^i, \tilde{Y}^i)$'s; s.t. $\tilde{S}^{(1)}_{max} \leq \tilde{S}^{(2)}_{max} \leq \ldots \leq \tilde{S}^{(B)}_{max}$

$$\mathrm{CI}_{1-\alpha} = [\tilde{S}^{(\lfloor \frac{\alpha}{2} B \rfloor)}_{max}, \tilde{S}^{(\lfloor (1-\frac{\alpha}{2}) B \rfloor)}_{max}]$$

Permutation p-value

i-th Permutation sample $\hat{X}^i$ (Y the same)



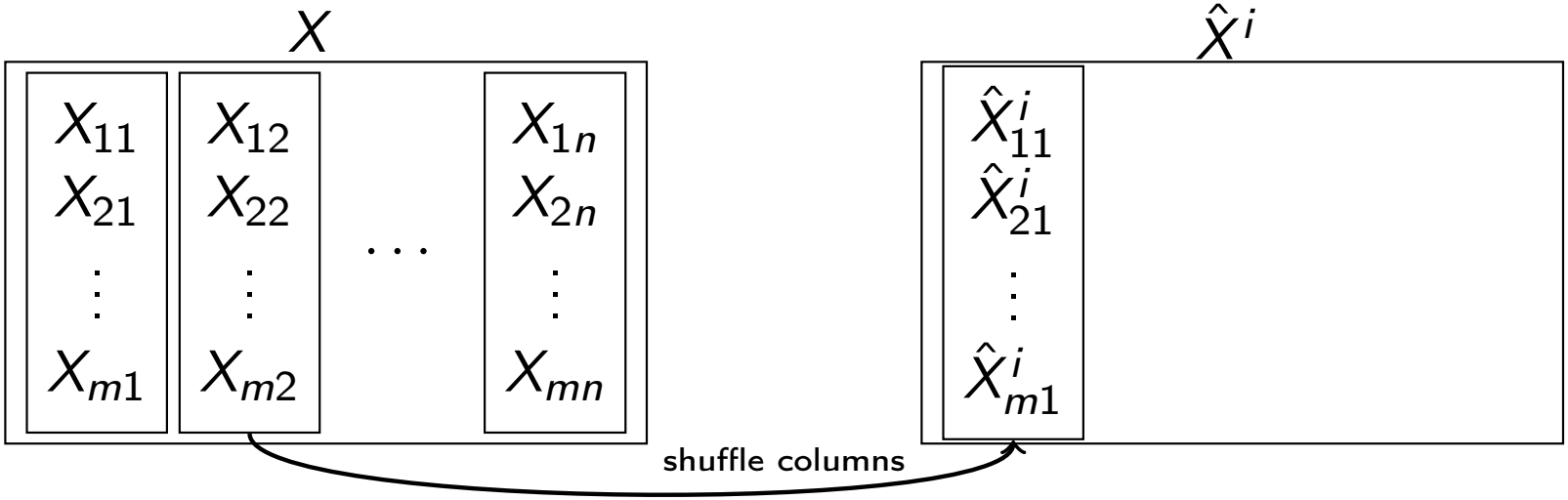
p-value

H_0: X and Y are two independent time series.

$$p-value = Prob[\hat{S} \geq S_{max}(X, Y)] \approx \frac{1}{L} \sum_{k=1}^{L} I[S_{max}(\hat{X}^k, Y) \geq S_{max}(X, Y)]$$

Permutation p-value

i-th Permutation sample $\hat{X}^i$ (Y the same)



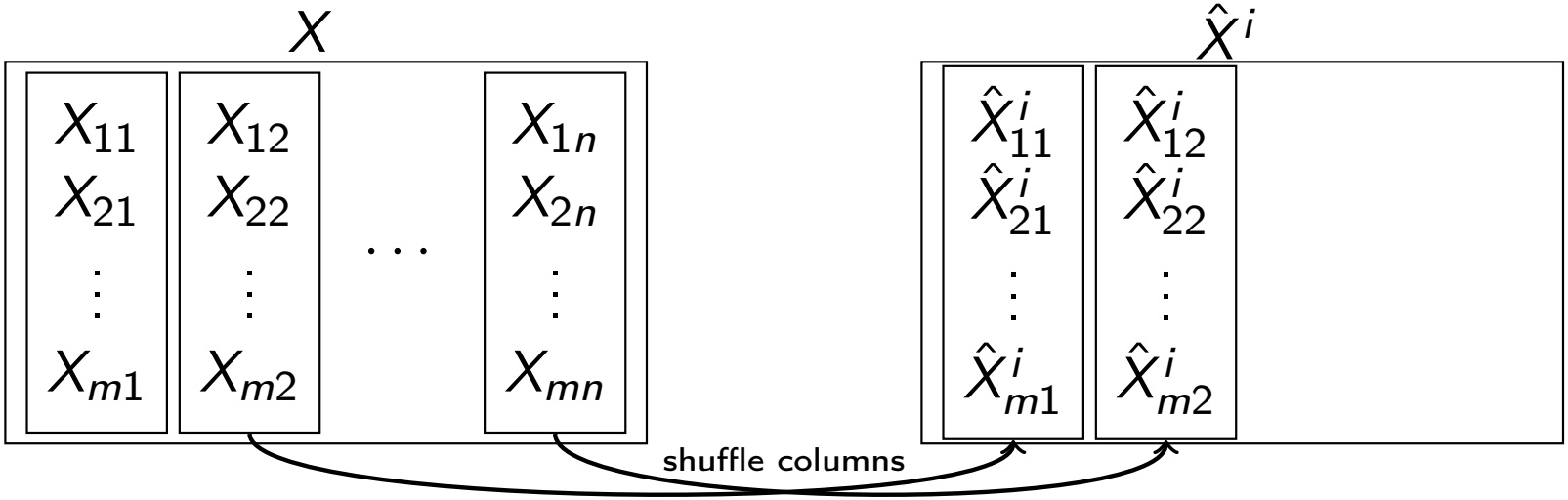
p-value

H_0: X and Y are two independent time series.

$$p-value = Prob[\hat{S} \geq S_{max}(X,Y)] \approx \frac{1}{L}\sum_{k=1}^{L} I[S_{max}(\hat{X}^k, Y) \geq S_{max}(X,Y)]$$

Permutation p-value

i-th Permutation sample $\hat{X}^i$ (Y the same)



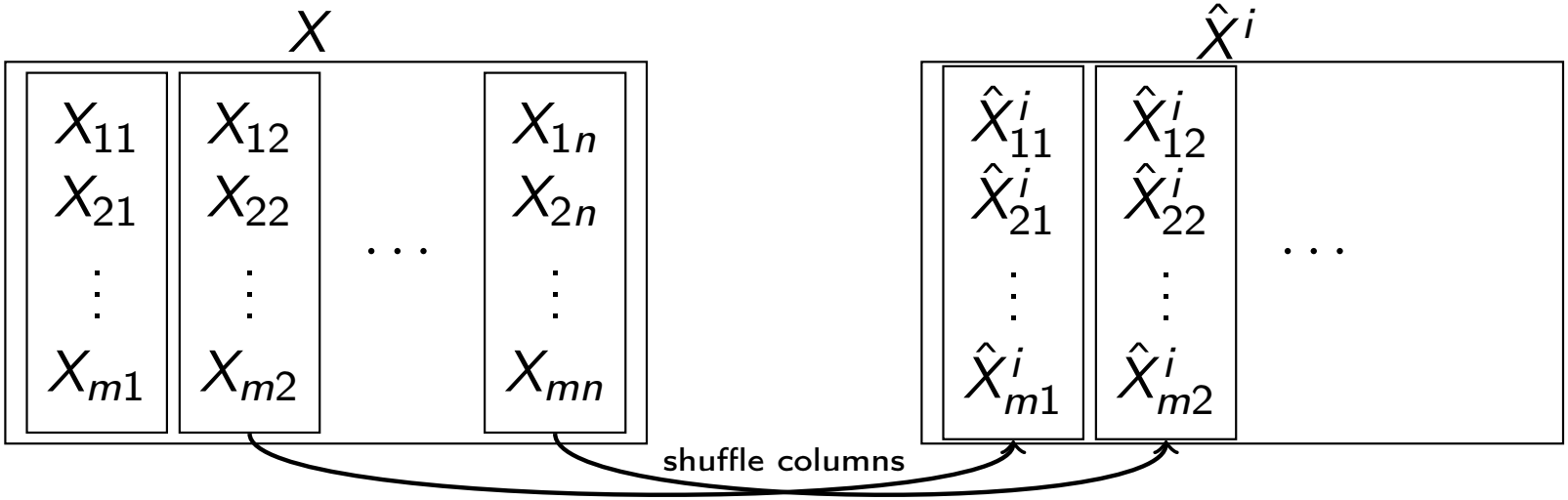
p-value

H_0: X and Y are two independent time series.

$$p-value = Prob[\hat{S} \geq S_{max}(X, Y)] \approx \frac{1}{L} \sum_{k=1}^{L} I[S_{max}(\hat{X}^k, Y) \geq S_{max}(X, Y)]$$

Permutation p-value

i-th Permutation sample $\hat{X}^i$ (Y the same)



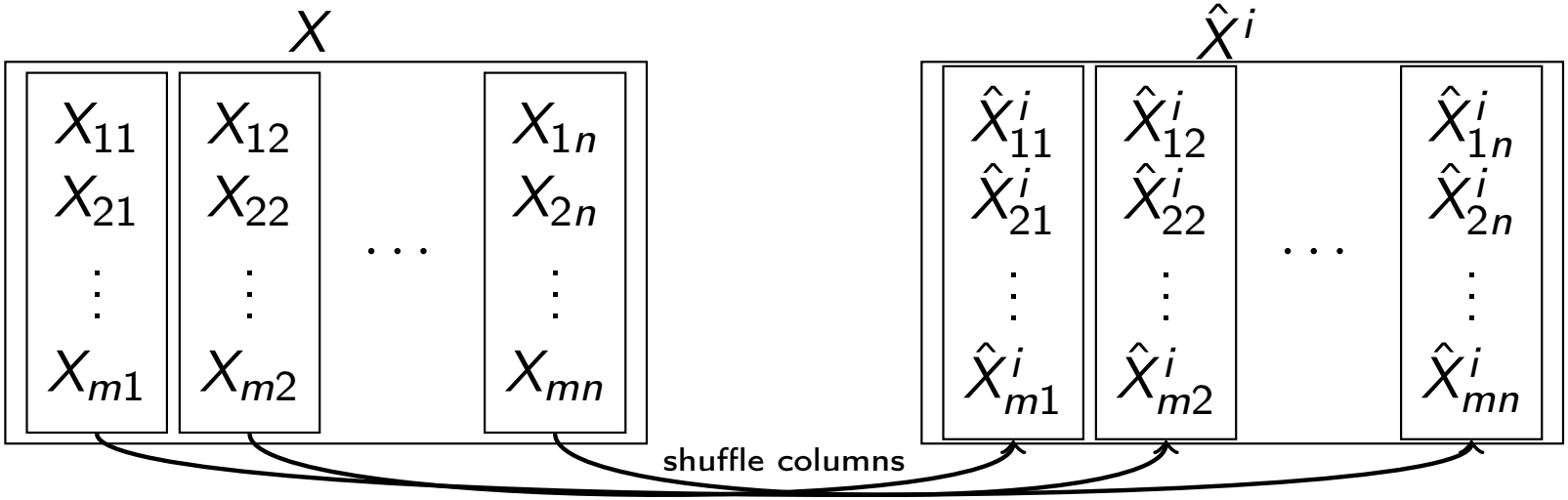
p-value

H_0: X and Y are two independent time series.

$$p-value = Prob[\hat{S} \geq S_{max}(X, Y)] \approx \frac{1}{L} \sum_{k=1}^{L} I[S_{max}(\hat{X}^k, Y) \geq S_{max}(X, Y)]$$

Permutation p-value

i-th Permutation sample $\hat{X}^i$ (Y the same)



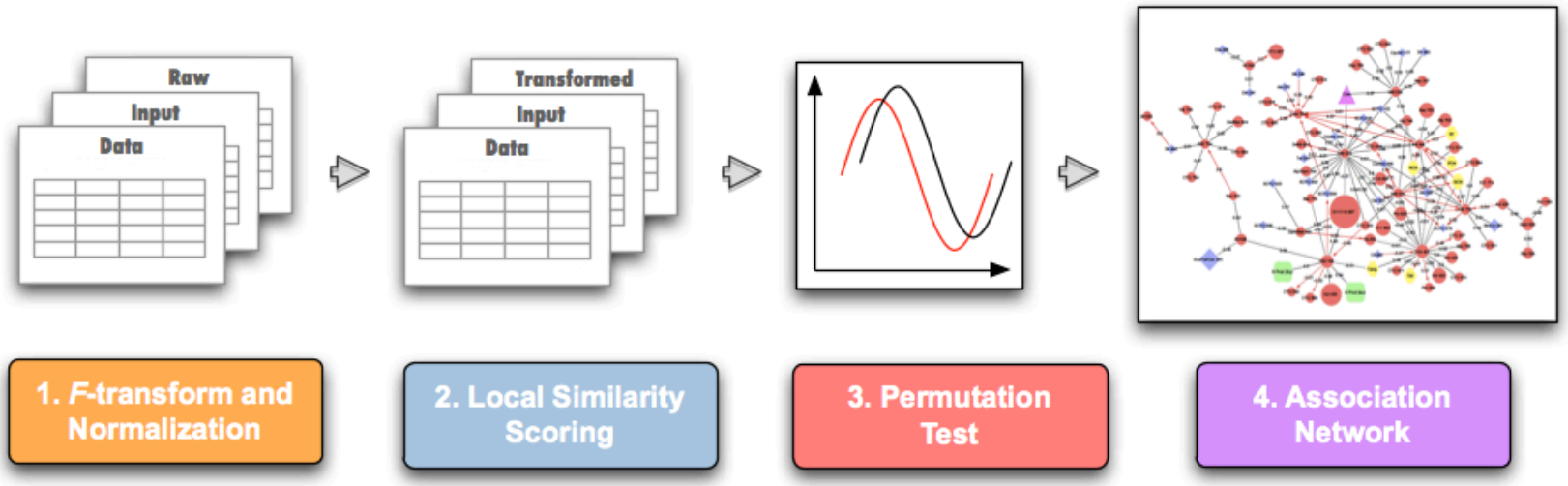
p-value

H_0: X and Y are two independent time series.

$$p-value = Prob[\hat{S} \geq S_{max}(X, Y)] \approx \frac{1}{L} \sum_{k=1}^{L} I[S_{max}(\hat{X}^k, Y) \geq S_{max}(X, Y)]$$

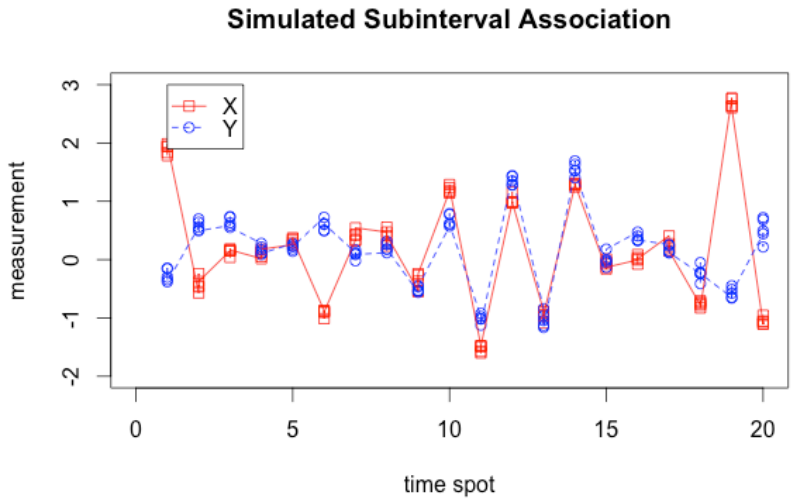
eLSA pipeline

Integrate into software pipeline



eLSA simulation

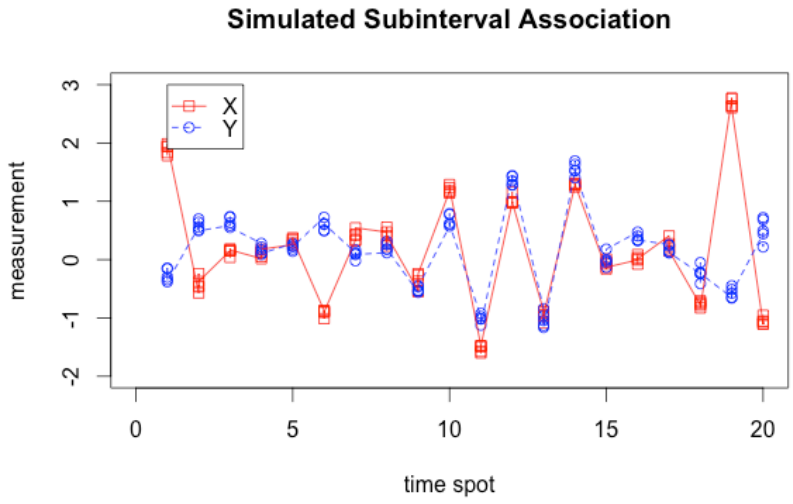
eLSA can identify time-dependent associations

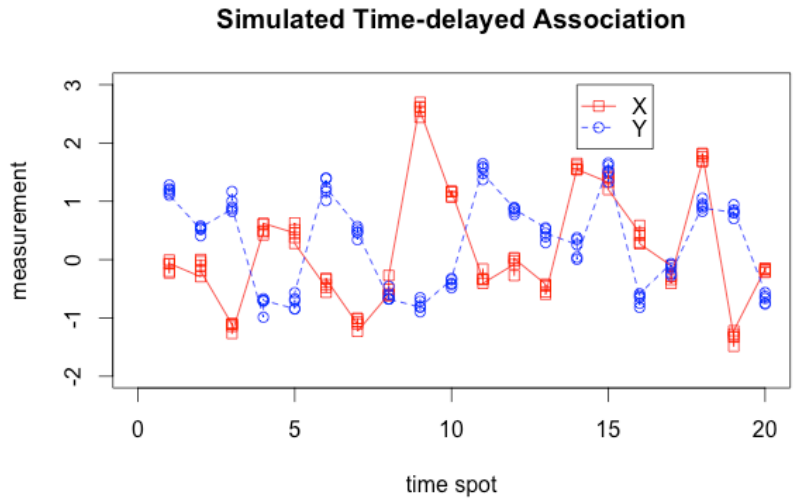


- PCC: r=0.26 (P=0.273)
- eLSA: LS=0.43 (P=0.028)
  $CI_{95\%}$=[0.43, 0.44]

eLSA simulation

eLSA can identify time-dependent associations





- PCC: r=0.26 (P=0.273)
- eLSA: LS=0.43 (P=0.028) $CI_{95\%}$=[0.43, 0.44]

- PCC: r=-0.26 (P=0.272)
- eLSA: LS=0.51 (P=0.006) $CI_{95\%}$=[0.50, 0.52]

Real Dataset Analysis

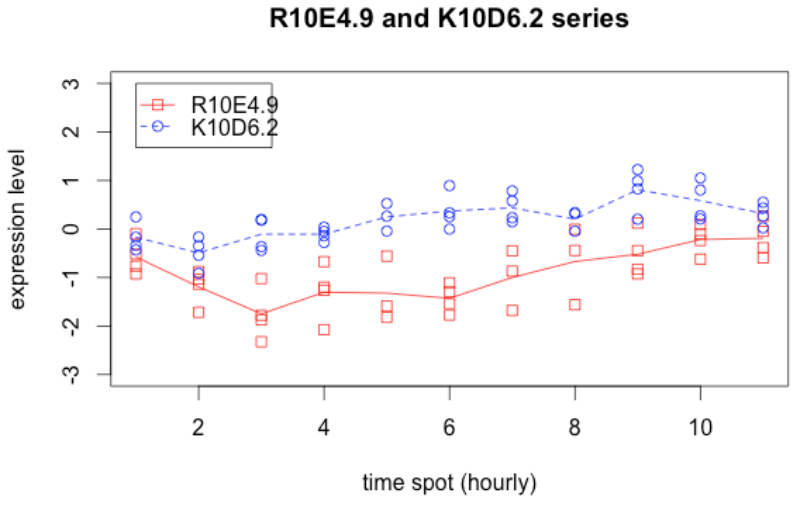
Real Datasets

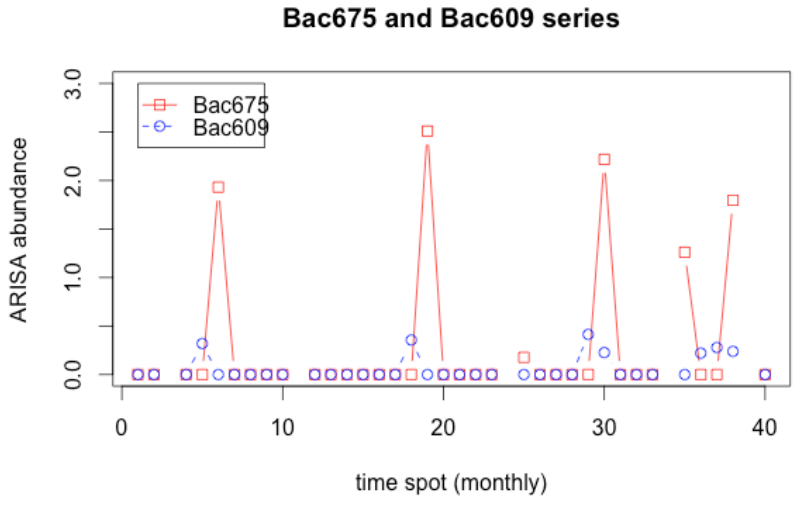
- *C. elegans*: 12 hours 4 replicates gene expression data *Wang and Kim (2003)*.
- Marine Microbial Community: coastal water near Los Angeles, 35 months non-replicated *Steele et al. (2011); Countway et al. (2010)*.

Summary

		$P \leq 0.01$ and $Q \leq 0.01$			$P \leq 0.05$ and $Q \leq 0.05$		
Dataset	# of factors	eLSA	PCC	both	eLSA	PCC	both
C. elegans	446	42532	56605	39114	57991	71799	54201
Microbial	515	1643	3237	293	2804	4242	658

Examples of Co-varying Associations

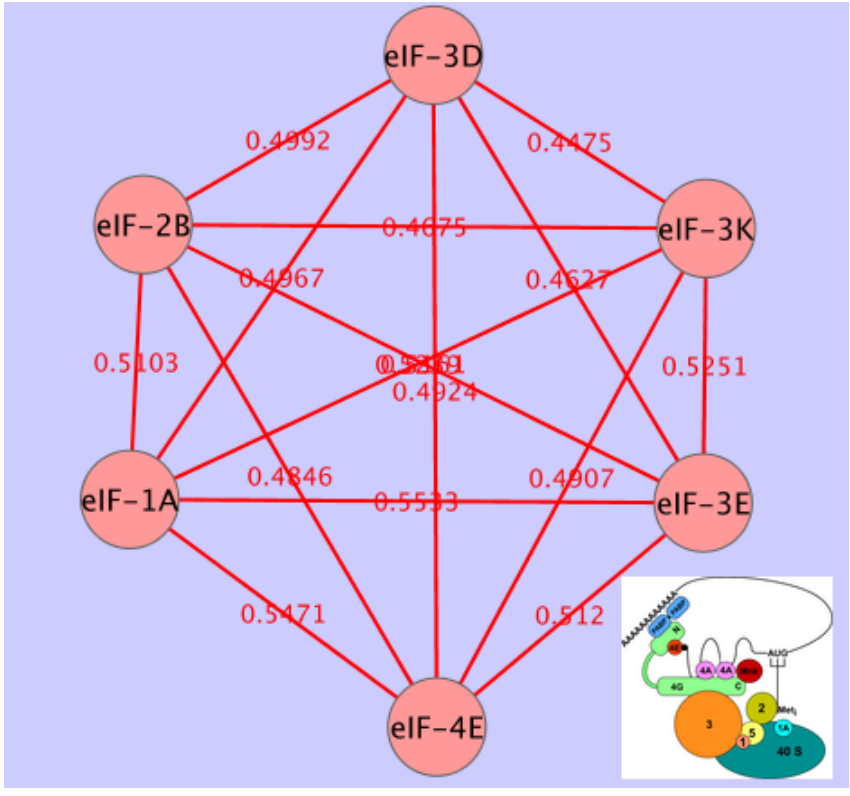




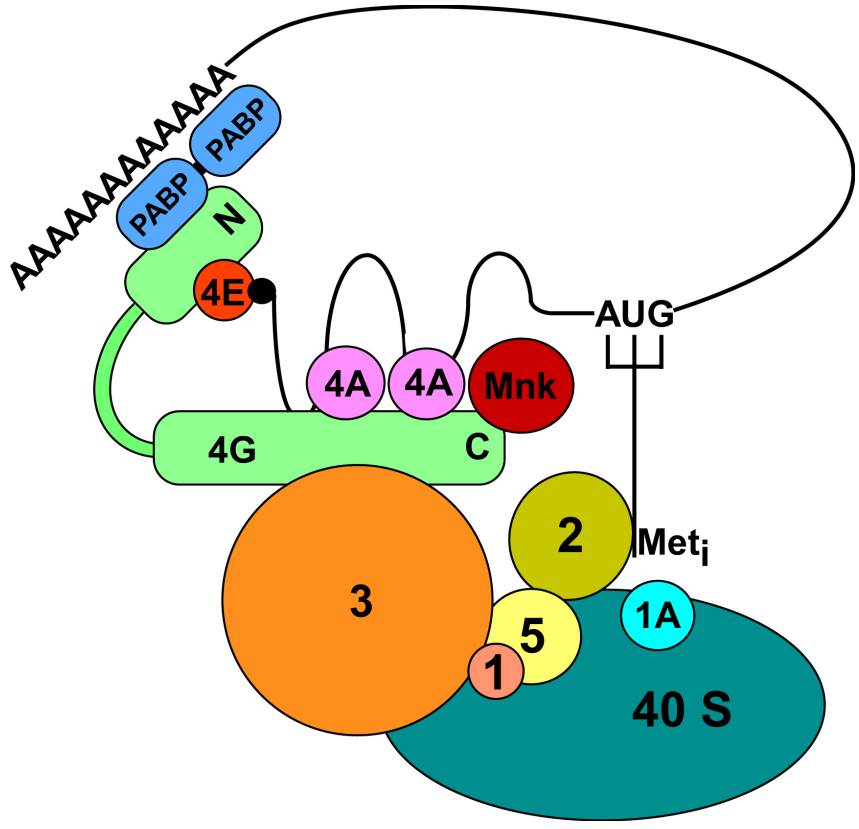
- *R*10*E*4.9 and *K*10*D*6.2 membrane protein genes
- *K*10*D*6.2 leads *R*10*E*4.9 one hour

- Show almost regular yearly pattern
- *Bac*609 leads *Bac*675 one month

Examples of Co-regulated Genes

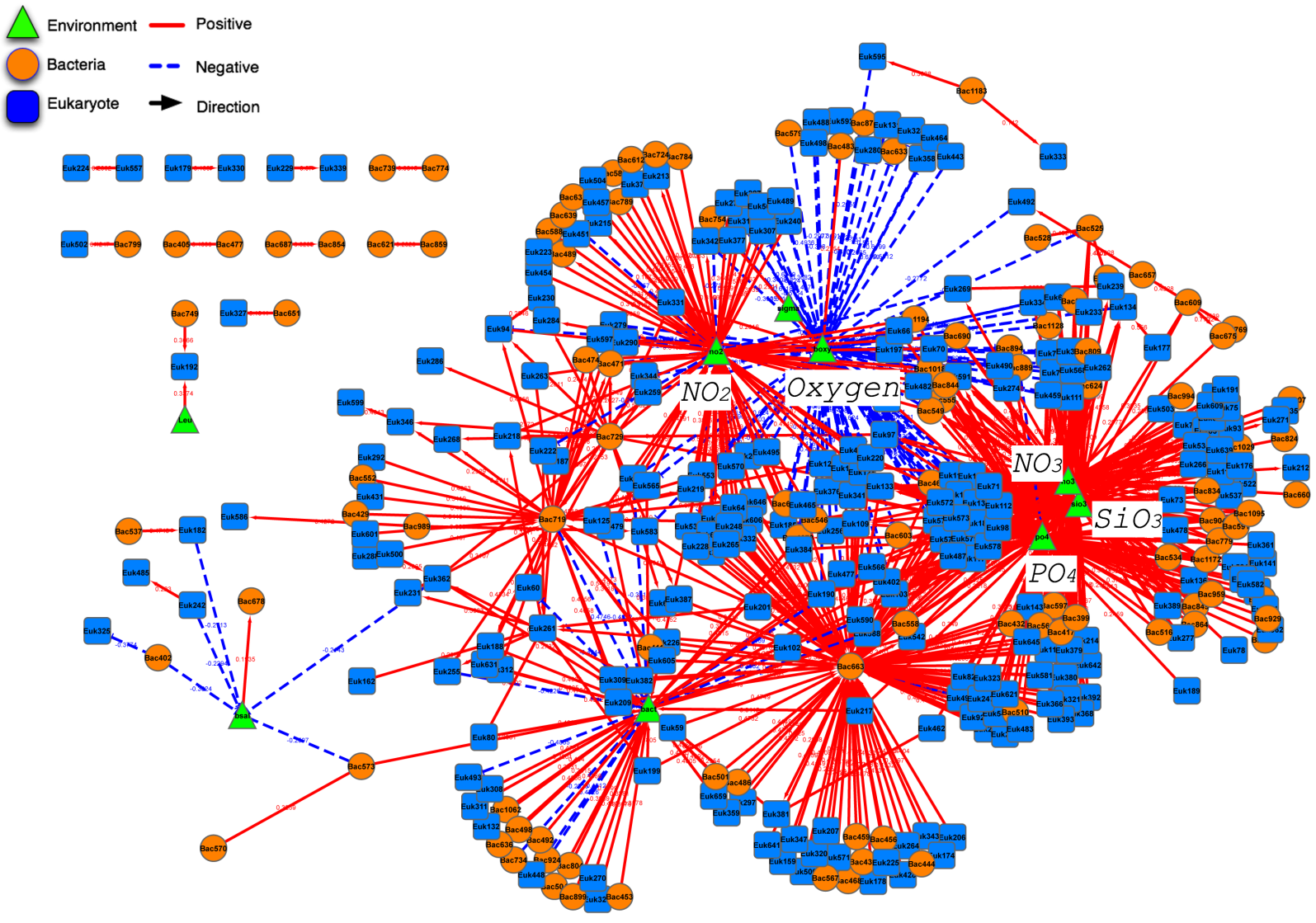


- eLSA inferred associations among elongation Initiations Factors (eIFs)



- eIF complex in *C. elegans* (figure from *Rhoads et al. (2006)*)

Association Network for Marine Data



Conclusions

Features of eLSA

- Captures local and possibly time-delayed associations.
- Choices of summarizing function for replicated data.
- Confidence Interval inference for LS score.
- Streamlines data normalization, local similarity scoring, permutation testing and network construction.
- Applicable to data with general gradients, including the response to different levels of treatments, temperature, humidity, or spatial distributions.

Availability

The eLSA software and web service are available from eLSA's homepage at `http://meta.usc.edu/softs/lsa`

Outline

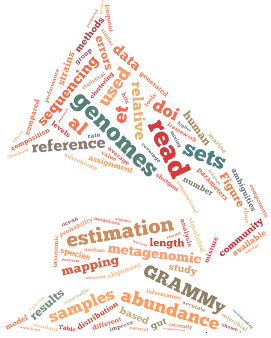
- Introduction
  - Biological Background
  - Our Computational Approaches
- GRAMMy for Shotgun Metagenomics
  - Motivations
  - Methods
  - Simulation Results
  - Application to Real Datasets
  - Conclusions
- eLSA for Time Series Data
  - Biological Background
  - Methods
  - Simulation Results
  - Application to Real Datasets
  - Conclusions
- Future work
- References

Thank you!

Thanks to my professors

Profs. Fengzhu Sun, Jed Fuhrman, Ting Chen and Jay Kuo.

Questions and comments



http://meta.usc.edu/softs/grammy

http://meta.usc.edu/softs/lsa

Outline

- Introduction
  - Biological Background
  - Our Computational Approaches
- GRAMMy for Shotgun Metagenomics
  - Motivations
  - Methods
  - Simulation Results
  - Application to Real Datasets
  - Conclusions
- eLSA for Time Series Data
  - Biological Background
  - Methods
  - Simulation Results
  - Application to Real Datasets
  - Conclusions
- Future work
- References

References I

Augenlicht, L. H. and Kobrin, D. (1982). Cloning and screening of sequences expressed in a mouse colon tumor. *Cancer Res*, 42(3):1088–93.

Avaniss-Aghajani, E., Jones, K., Chapman, D., and Brunk, C. (1994). A molecular technique for identification of bacteria using small subunit ribosomal RNA sequences. *Biotechniques*, 17(1):144–6, 148–9.

Countway, P., Vigil, P., Schnetzer, A., Moorthi, S., and Caron, D. (2010). Seasonal analysis of protistan community structure and diversity at the USC Microbial Observatory (San Pedro Channel, North Pacific Ocean). *Limnol Oceanogr*, 55(6):2381–2396.

DeLong, E. F. (2005). Microbial community genomics in the ocean. *Nat Rev Microbiol*, 3(6):459–69.

References II

Fisher, M. M. and Triplett, E. W. (1999). Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. *Appl Environ Microbiol*, 65(10):4630–6.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol*, 5(10):R245–9.

Hao, X., Jiang, R., and Chen, T. (2011). Clustering 16s rrna for otu prediction: a method of unsupervised bayesian clustering. *Bioinformatics*, 27(5):611–8.

Paver, S. F. and Kent, A. D. (2010). Temporal patterns in glycolate-utilizing bacterial community composition correlate with phytoplankton population dynamics in humic lakes. *Microb Ecol*, 60(2):406–18.

References III

Rhoads, R. E., Dinkova, T. D., and Korneeva, N. L. (2006). Mechanism and regulation of translation in c. elegans. *WormBook : the online review of C. elegans biology*, pages 1–18.

Ruan, Q., Dutta, D., Schwalbach, M. S., Steele, J. A., Fuhrman, J. A., and Sun, F. (2006a). Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics*, 22(20):2532–8.

Ruan, Q., Steele, J. A., Schwalbach, M. S., Fuhrman, J. A., and Sun, F. (2006b). A dynamic programming algorithm for binning microbial community profiles. *Bioinformatics*, 22(12):1508–14.

Shade, A., Chiu, C. Y., and McMahon, K. D. (2010). Differential bacterial dynamics promote emergent community robustness to lake mixing: an epilimnion to hypolimnion transplant experiment. *Environ Microbiol*, 12(2):455–66.

References IV

Sogin, M. L., Morrison, H. G., Huber, J. A., Mark Welch, D., Huse, S. M., Neal, P. R., Arrieta, J. M., and Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U S A*, 103(32):12115–20.

Steele, J. A., Countway, P. D., Xia, L., Vigil, P. D., Beman, J. M., Kim, D. Y., Chow, C. E., Sachdeva, R., Jones, A. C., Schwalbach, M. S., et al. (2011). Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *ISME J.*

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–5.

References V

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., et al. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304(5667):66–74.

Wang, J. and Kim, S. K. (2003). Global analysis of dauer gene expression in caenorhabditis elegans. *Development*, 130(8):1621–34.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63.

Waterman, M. S. (1995). *Introduction to computational biology : maps, sequences and genomes*. Chapman & Hall, London ; New York, 1st edition.

Xia, L. C., Cram, J. A., Chen, T., Fuhrman, J. A., and Sun, F. (2011a). Accurate genome relative abundance estimation based on shotgun metagenomic reads. *PLoS One*, accepted.

References VI

Xia, L. C., Steele, J. A., Cram, J. A., Cardon, Z. G., Simmons, S. L., Vallino, J. J., Fuhrman, J. A., and Sun, F. (2011b). Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. *BMC Systems Biology*, 5(Suppl 2):S15.

Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., et al. (2001). Global analysis of protein activities using proteome chips. *Science*, 293(5537):2101–2105.