# Large Scale Prediction of Transcription Factor Binding Sites for Gene Regulation using Cloud Computing

Zhengchang Su

Department of Bioinformatics and Genomics
The University of North Carolina at Charlotte

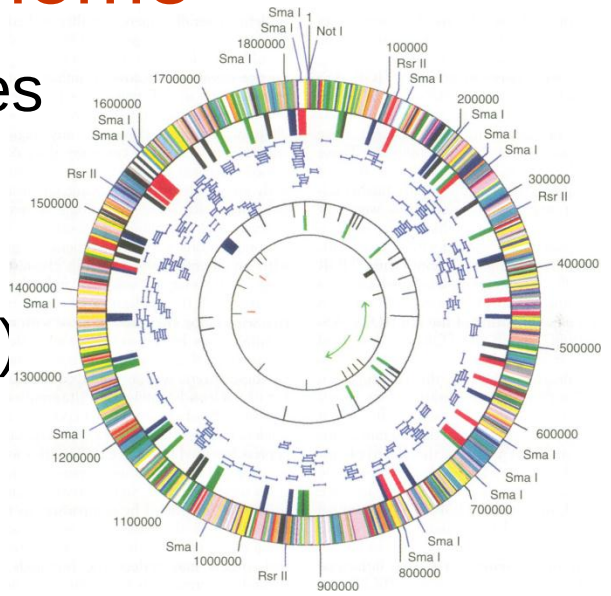
6-2-2011 at Cloud Futures Workshop 2011

# Outline

- Introduction to the problem
- GleClubs: an algorithm for large scale prediction of transcription factor binding sites (TFBSs)
- Parallelization of GleClubs on a distributed memory cluster using MPI
- Porting GleClubs on Windows Azure Platform

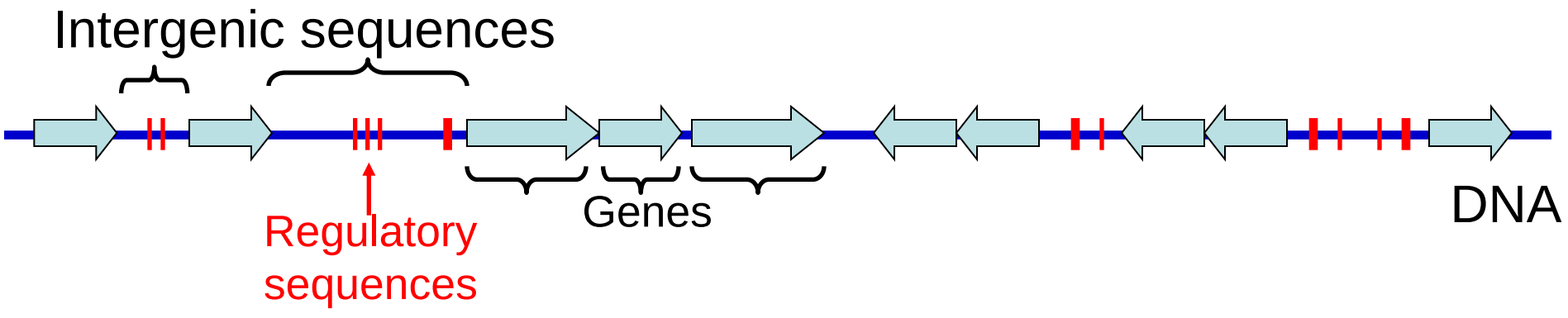
# Structure of a bacterial genome

- There are two types of functional sequences in a bacterial genome:



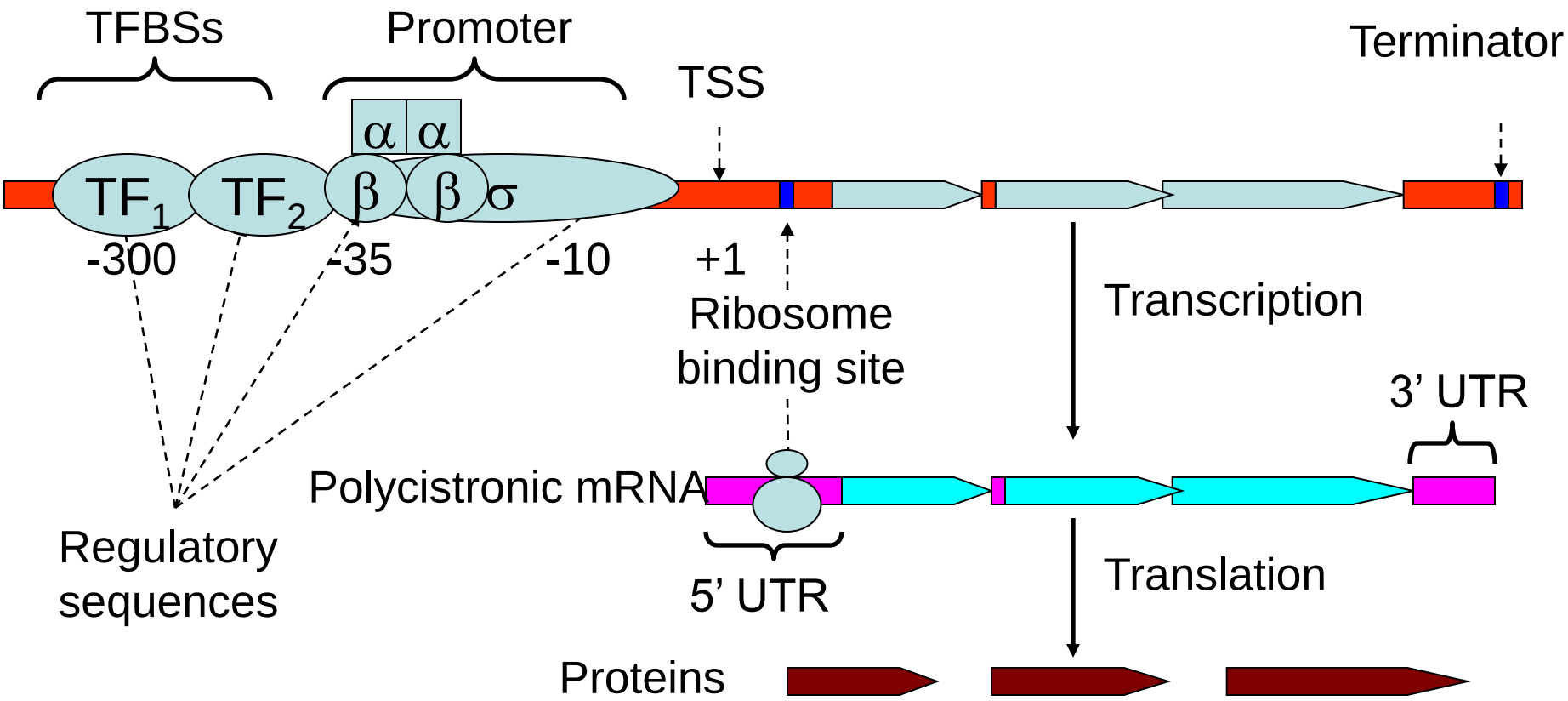

1. **Coding sequences/genes:** specify the cellular components (proteins and RNAs) of an organism; consist of ~85% of genome sequences.

2. **Regulatory sequences:** specify when, where, how fast, and how much the product of a gene should be produced; usually are located in intergenic sequences, consisting of~15% of genome sequences.



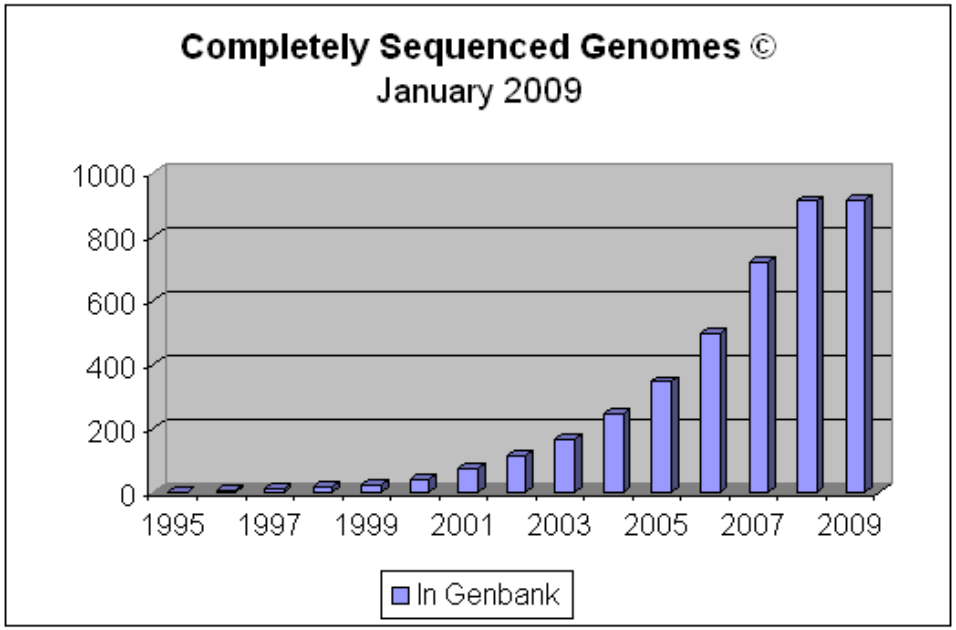
# The functions of regulatory sequences

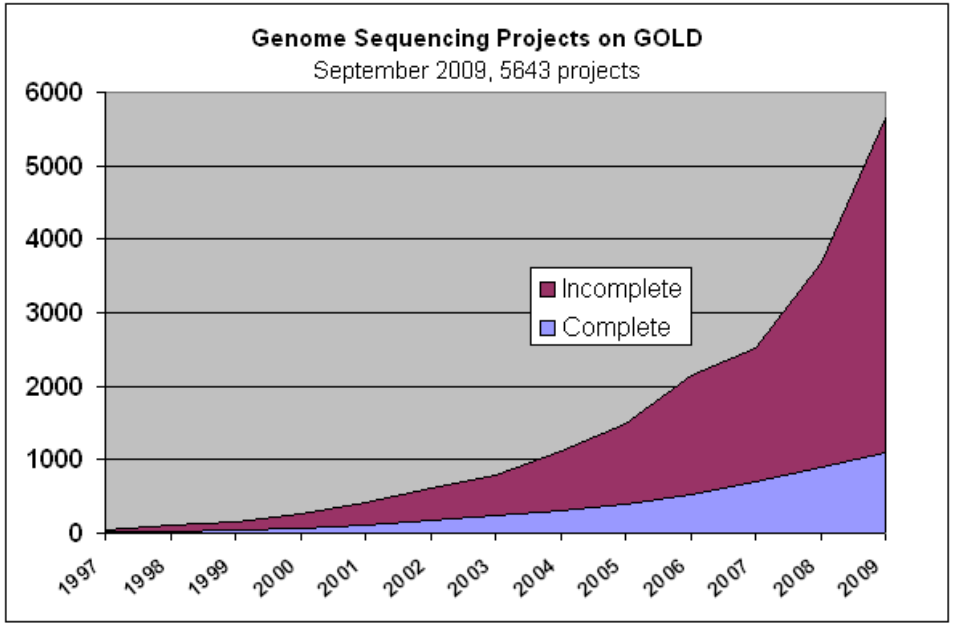
- Regulatory sequences control gene expression through interacting with a transcription factor (TF), thus, they are also called **TF binding sites (TFBSs)**.
- Adjacent genes in a bacterial genome are often organized in a transcription unit called an **operon**.



# Explosion of genome sequence data

- Since 1995, and particularly, since 2007 the number of sequenced genomes increases exponentially.




As of 5-28-2011 http://www.genomesonline.org

| | Archaea | Bacteria | Eukaryota | Total |
|---|---|---|---|---|
| Complete | 109 | 1486 | 154 | 1749 |
| In pipline | 201 | 6045 | 2006 | 8252 |

# A paramount goal of computational biology

- One of the most challenging goals of computational biology is to understand the functions of an organism solely from its genome sequences.
- The first step towards this goal is to know the part list of its cells:
  - **Genes:** encoding cellular components for biological functions

    **The gene finding problem:** is relatively well-solved, so we know vast majority of genes in almost any sequenced prokaryotic genomes—more than 1,486 genomes.

  - **TFBSS**, encoding controlling programs for making cellular components

    **The regulatory sequence finding problem:** remains a largely unsolved-problem, so we know very little about TFBSs in almost all sequenced genomes.

# Causes for the difficulty of *cis*-regulatory sequence prediction

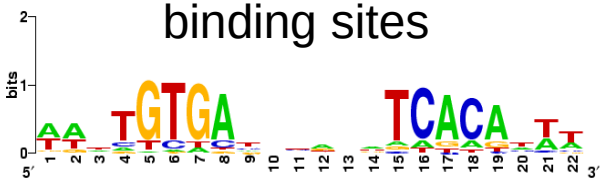
- *cis*-regulatory elements are short (6-25bs) and degenerate;
- They are located in intergenic regions that are much longer;
- There is usually no base usage bias, any sequence segment can be potentially a *cis*-regulatory element.

Examples of $\sigma^{70}$ binding sites in the *E. coli* genome

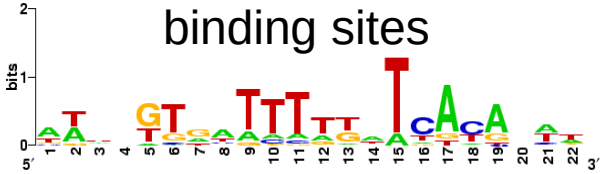
**TACGAT**
**TATAAT**
**TATAAT**
**GATACT**
**TATGAT**
**TATGTT**

Examples of CRP binding sites in the *E. coli* genome

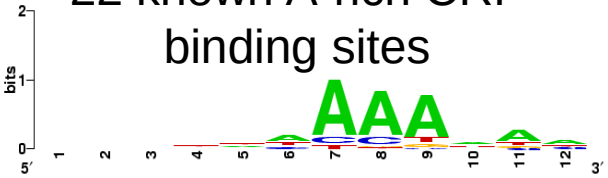
150 known canonical CRP binding sites



26 known T-rich CRP binding sites



22 known A-rich CRP binding sites



- A collection of similar binding sites of a TF is called a motif.

# The motif-finding problem

- Since there are usually no fixed patterns of *cis*-regulatory elements of a TF, a *cis*-regulatory element can only be predicted by comparing a set of sequences that are likely to contain the binding site of the same TF.

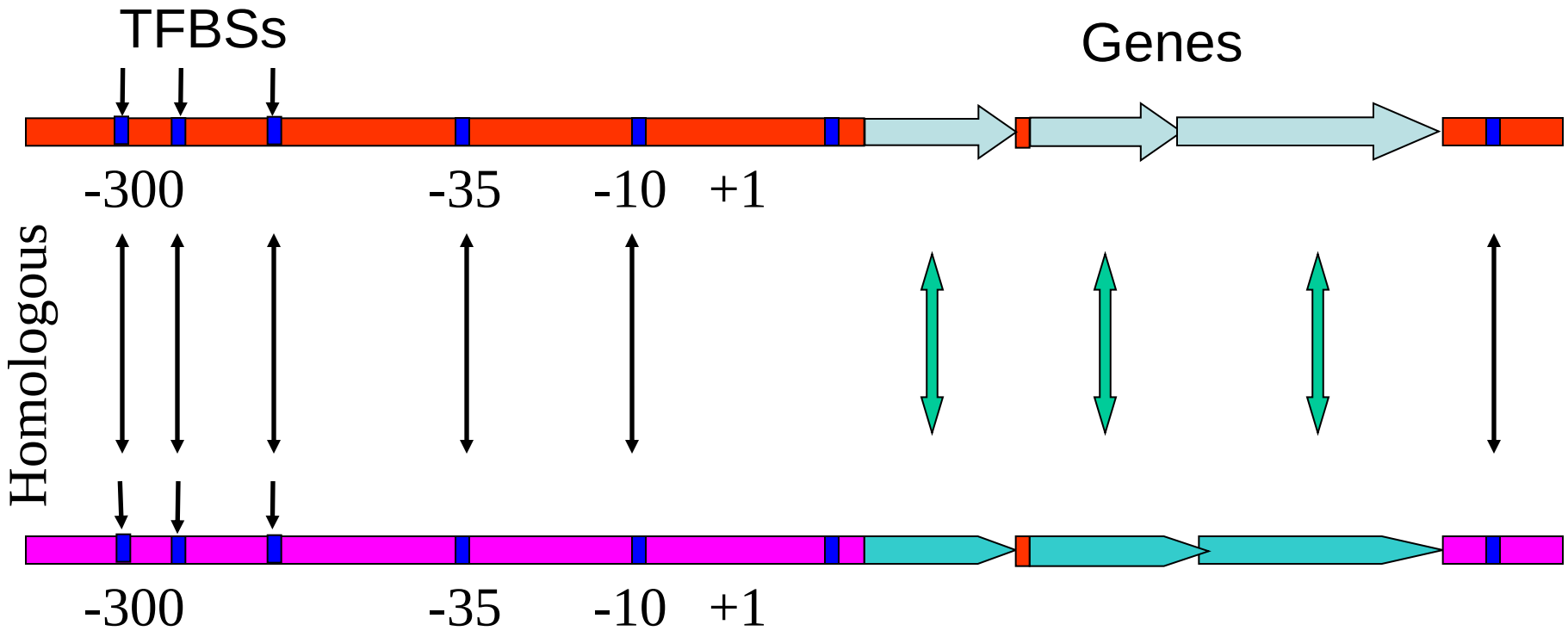
- The problem of finding *cis*-regulatory elements in a given set of sequences is called the **motif-finding problem**.

- Numerous sequence-based motif-finding algorithms and tools has been developed, and they are all based on the assumption that binding sites of a TF are more conserved than the flanking sequences.

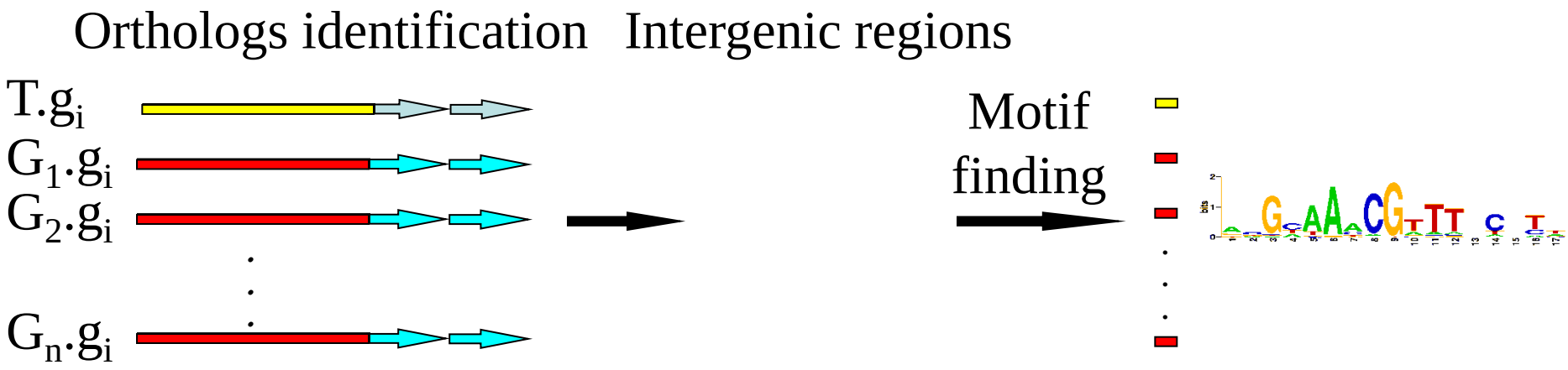
- Unfortunately, current motif-finding tools often return too many false positives while still missing true binding sites.

# Methods for finding a set of intergenic sequences for motif-finding

- **One gene, multiple genomes approach---phylogenetic footprinting:** in closely related species, more often both the coding sequences and regulatory sequences of orthologous genes are conserved.



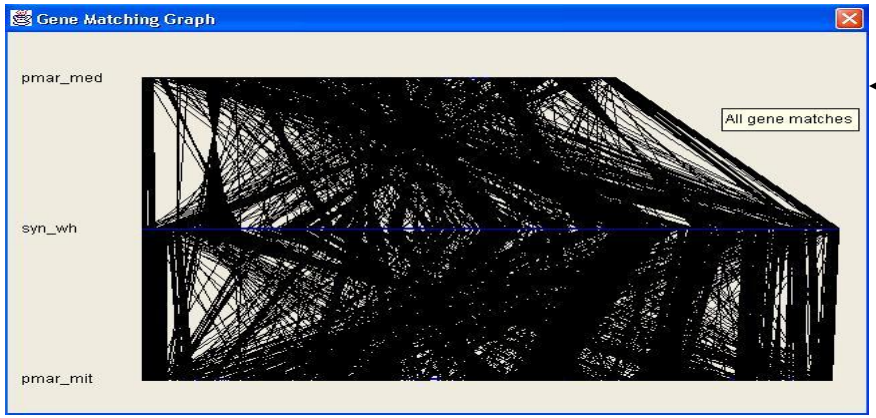
# A typical phylogenetic footprinting procedure




- This procedure can be applied to each gene in the target genome to predict TFBSs for all genes.
- Similar putative TFBSs are then clustered to form a motif.
- However, the prediction sensitivity and specificity of such an approach is very low due to the low prediction accuracy of motif-finding tools.
- The predictions are biased to the target genome, so the binding sites in reference genomes cannot be fully predicted.

# Prediction of operons

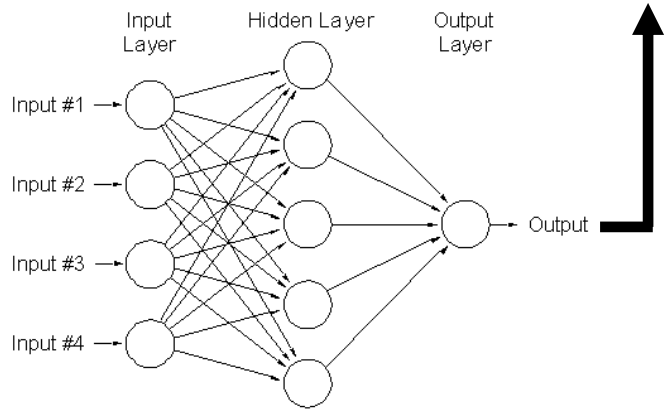
- An algorithm for operon prediction: JPOP(joint prediction of operons)



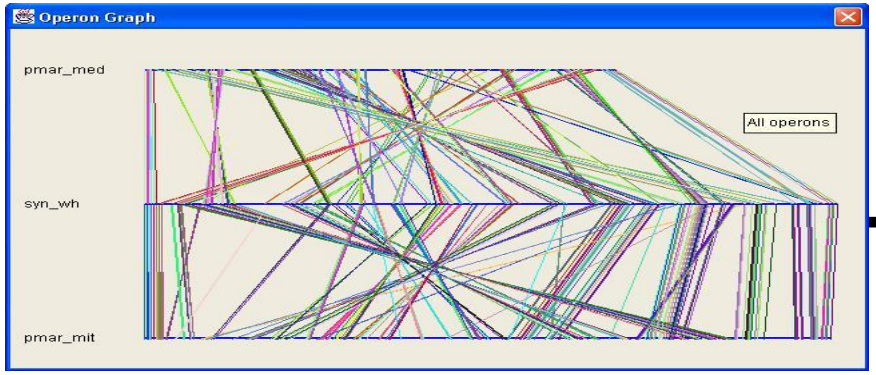

Homology mapping across multiple related genomes

Gene pairs on the same strand
- Operon pair
- Non-operon pair

1. intergenic distances
2. predicted gene functions (COG)
3. phylogenetic profiles
4. k-mer frequency statistics in the intergenic region



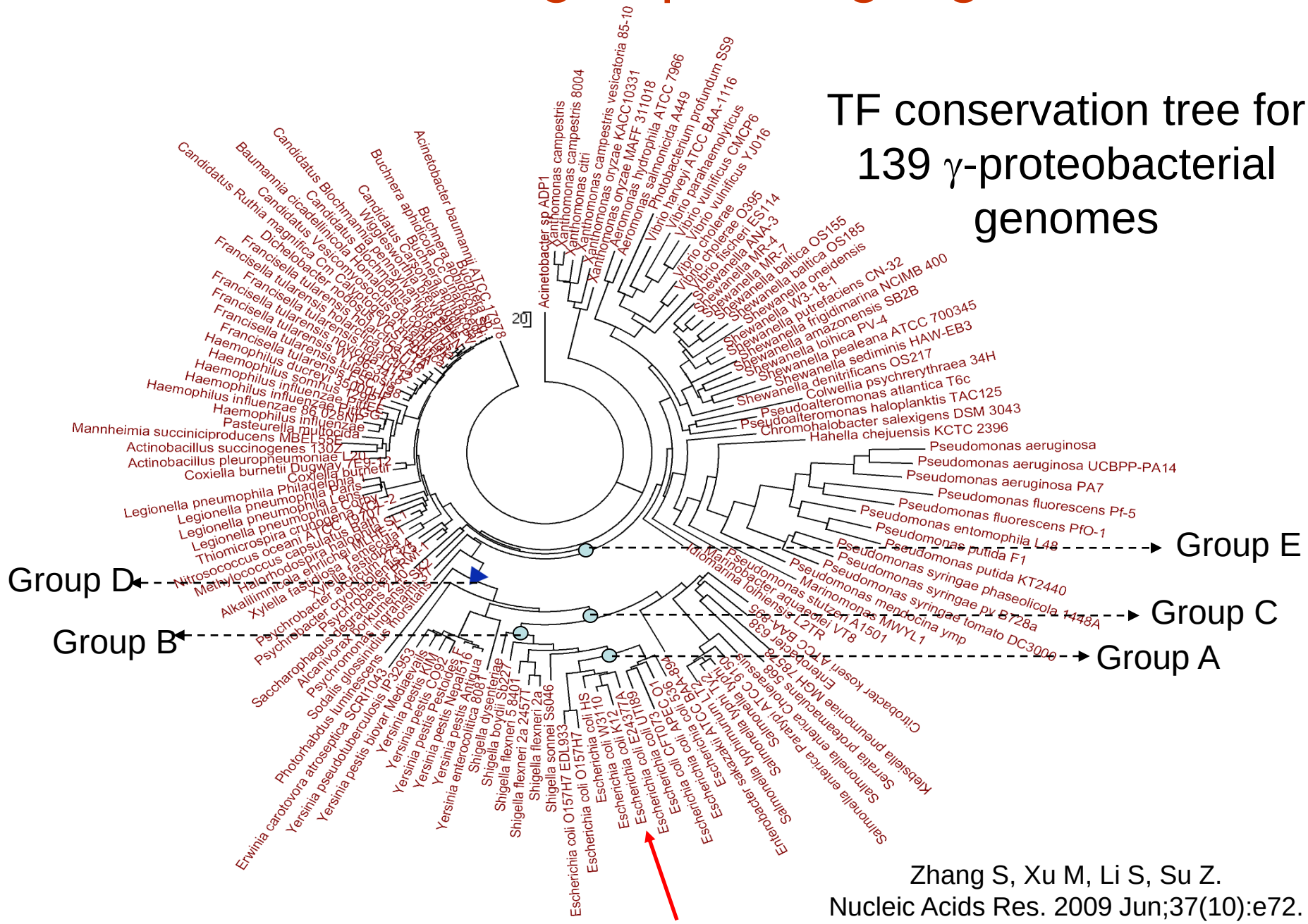

A neural network trained on *E. coli* data set




Conserved gene neighborhoods

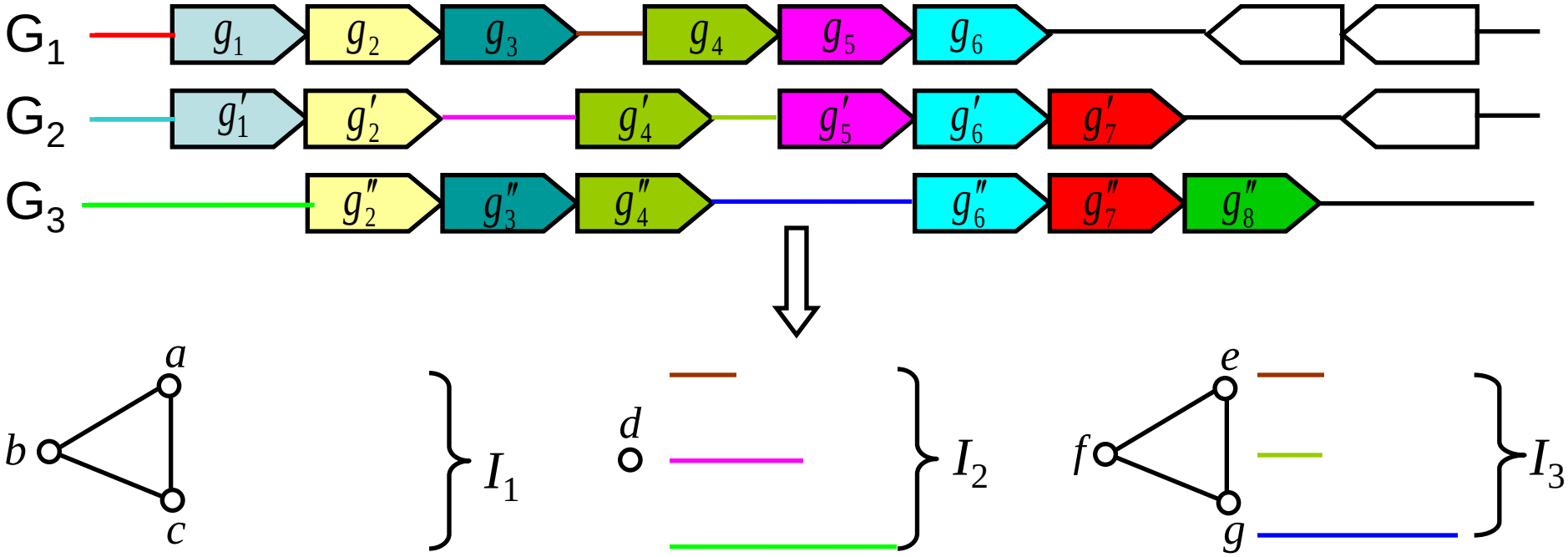
X. Chen, Z. Su, et al. NAR, 2004, 32(7):2147-57
P. Dam, V. Olman, K. Harris, Z. Su, Y. Xu. NAR. 2007;35(1):288-98.

# Selection of a group of target genomes



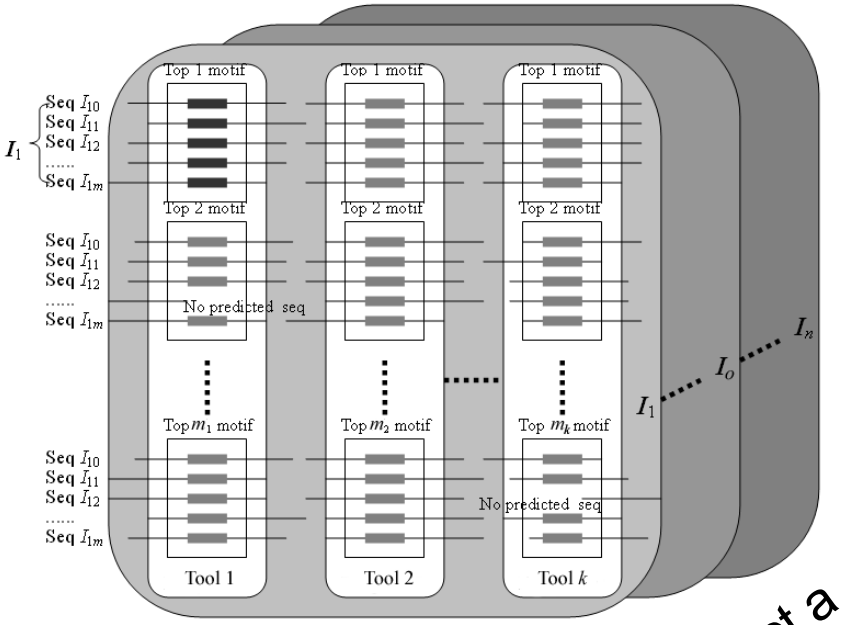

Zhang S, Xu M, Li S, Su Z.
Nucleic Acids Res. 2009 Jun;37(10):e72.

# Extraction of inter-operonic sequences associated with each orthologous operon groups




- Using a group of 32 γ-proteobacteria genomes (Group D) including *E. coli* K12, we identified 4,103 orthologous operon groups and the same number of inter-operonic sequence sets.

# Identifying dense subgraphs as possible motifs

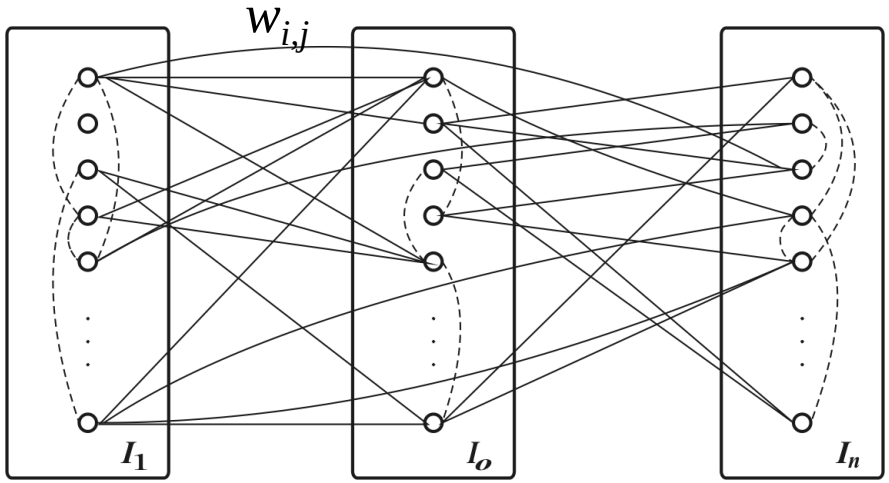



Input motifs

Construct a weighted graph

4,103 inter-operonic sequence sets x 40 motifs
= 160 x $10^3$ nodes

1.6 x $10^5$ x 1.6 x $10^5$ /2 = 13 x $10^9$ possible edges

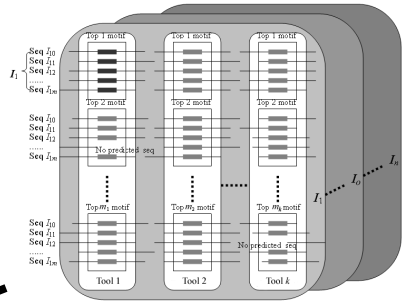



Motif similarity graph

- Two assumptions:
  1. A true motif is more likely to be identified by different tools in the same inter-operonic sequence set;
  2. A true motif is more likely to identified by in different sets of inter-operonic sequences sets.

# Flowchart of the motif clustering algorithm

Zhang S, Li S, Pham PT, Su Z.
BMC Bioinformatics. 2010;11:397



Input motifs

β

α

$G_2$ with a lower density

$G_1$ with a higher density

$G_2$

MCL

$c_1$ $c_2$ $c_3$ $c_4$ ⋮ $c_t$

Mapping the clusters to $G_1$

$G_1$ $c_2$ $c_3$ $c$ $c_4$ $c_t$ $c_1$

$G_3$ $c_3$ $c_2$ $c_4$ $c_t$ $c_1$

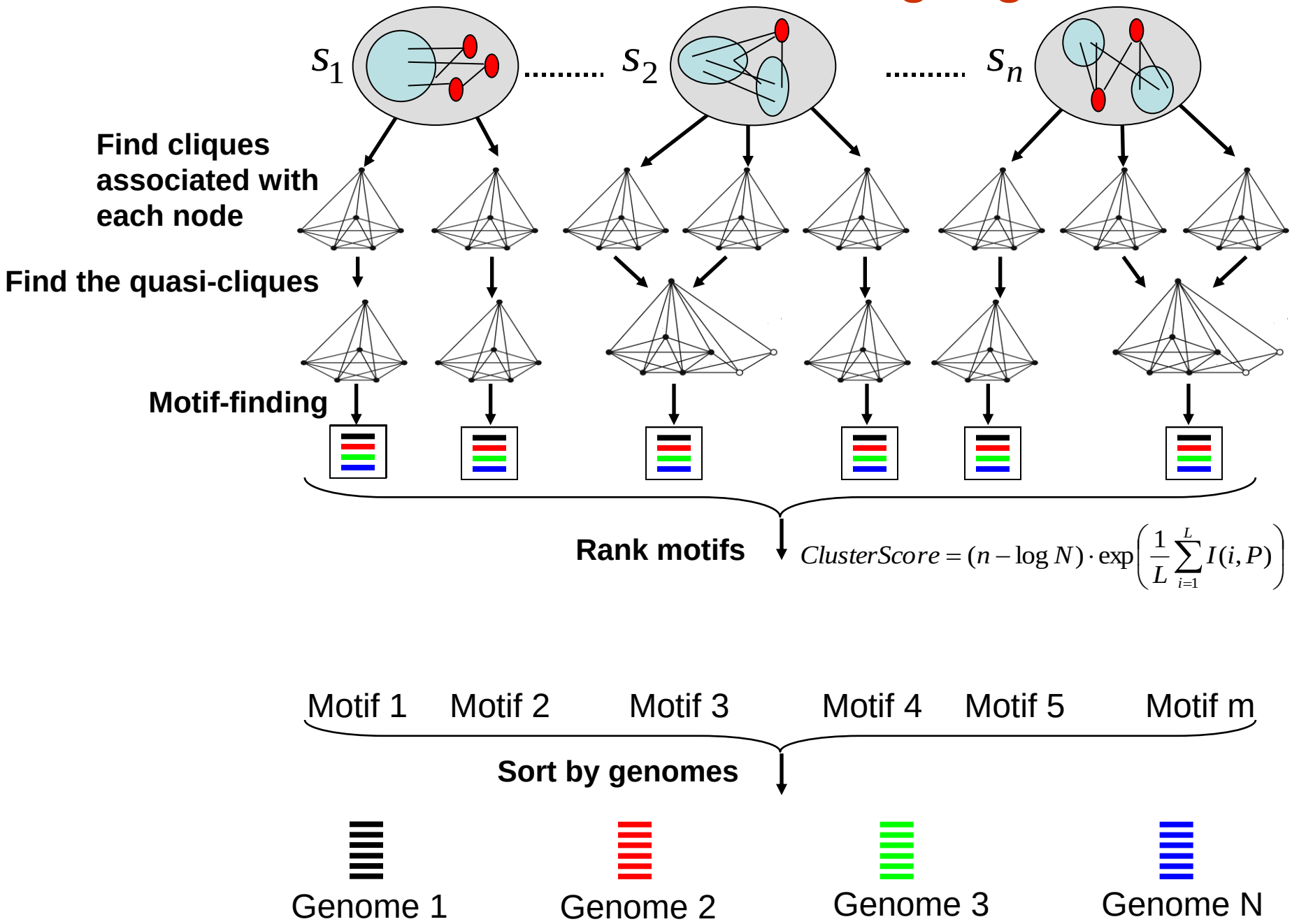
MCL

… 

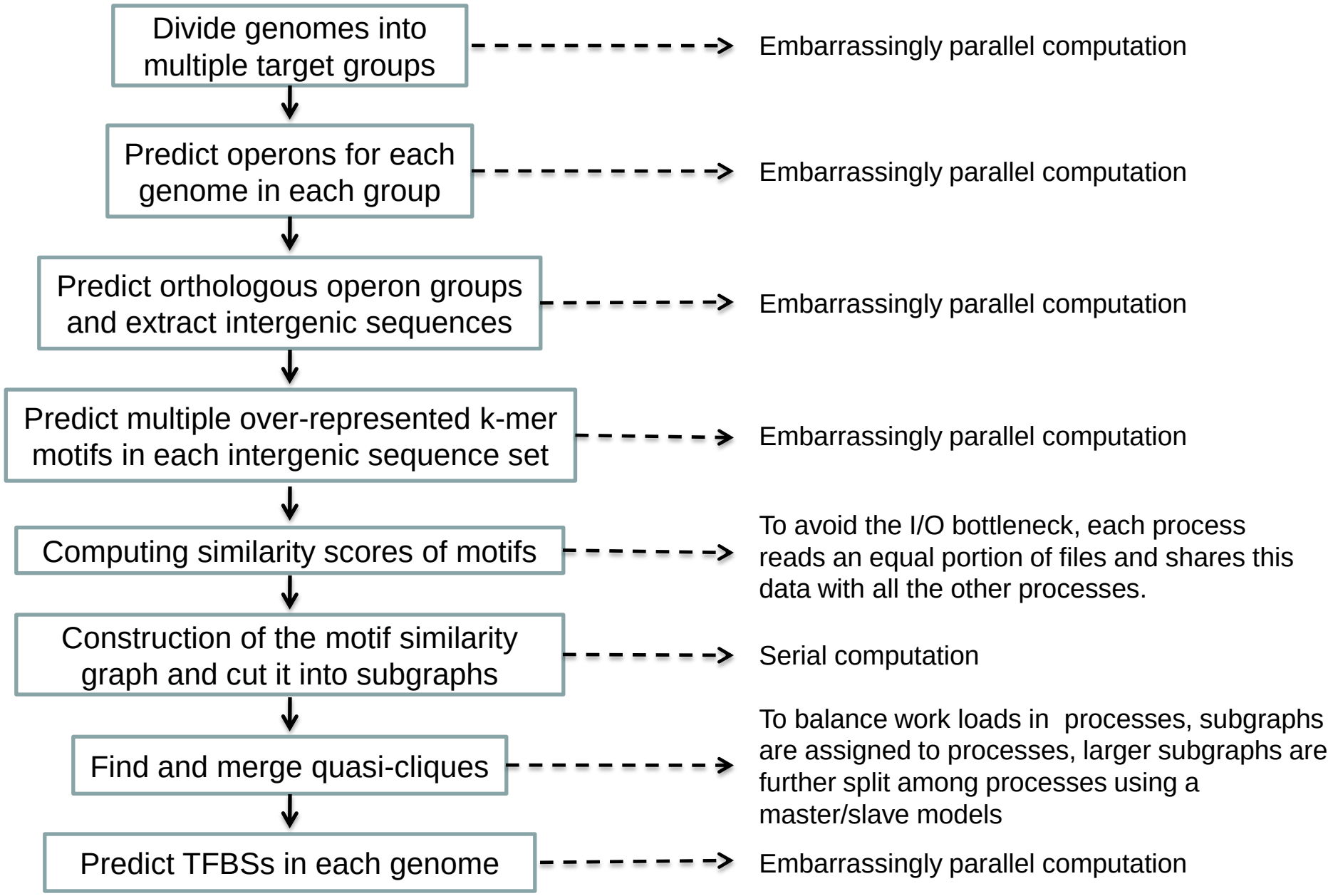
Inducing subgraphs in $G_1$

$S_1$ ..... $S_2$ ..... $S_n$

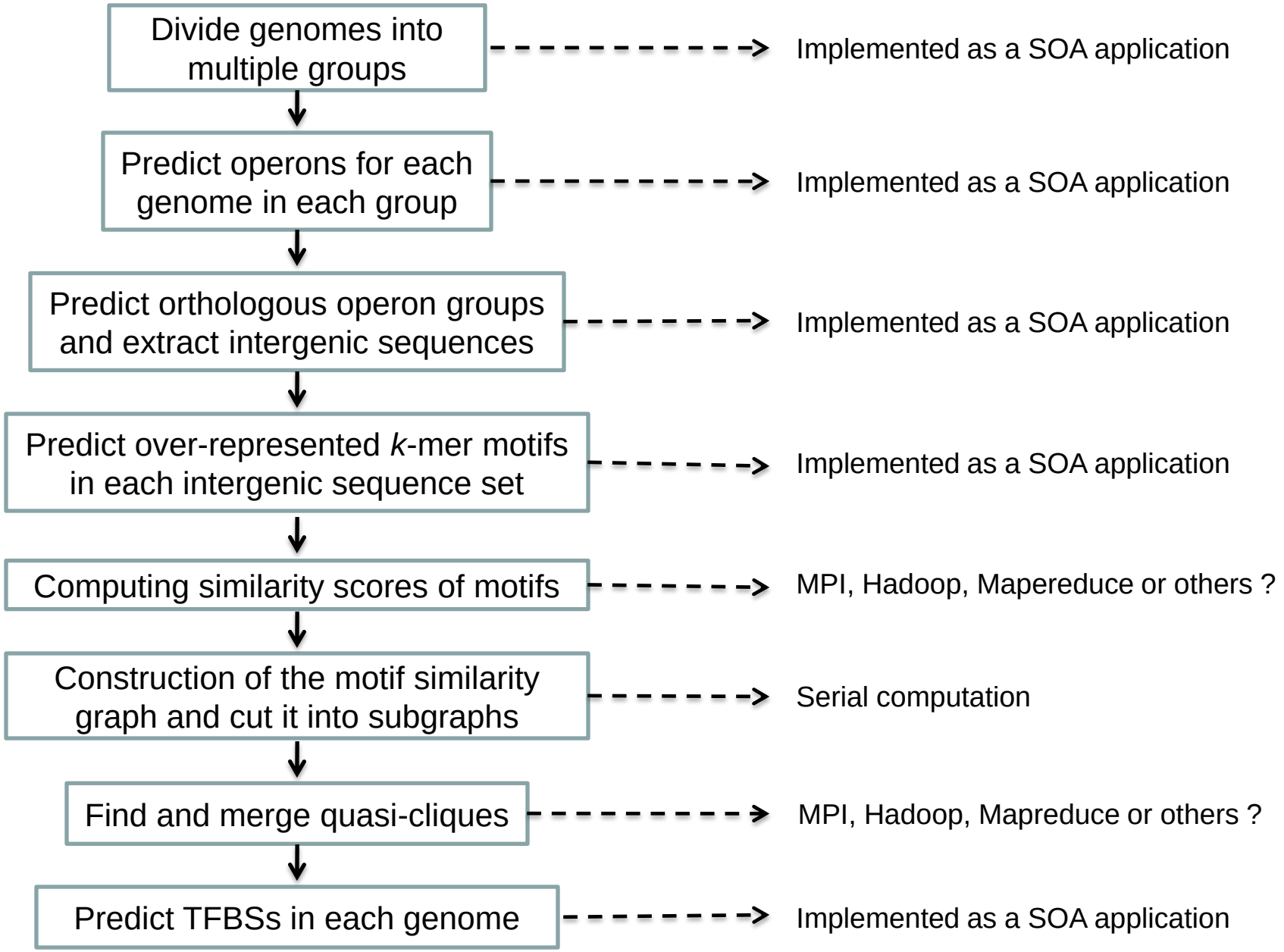
# Flowchart of the motif clustering algorithm



# Parallelization of GleClubs on a cluster



# Port GleClubs on the Windows Azure Platform



# Acknowledgments


- Dr. Shaoqiang Zhang
- Dr. Xia Dong
- Peter Pham
- Shan Li
- Meng Niu
- Minli Xu
- Chen Xu
- Matthew Smith

UNC Charlotte

- Dr. Barry Wilkinson
- Ridhi Dua
- Dr. Srinivas Akella
- Youjie Zhou

Funding sources:


National Science Foundation
WHERE DISCOVERIES BEGIN

Microsoft®

Thank You !

Questions ?