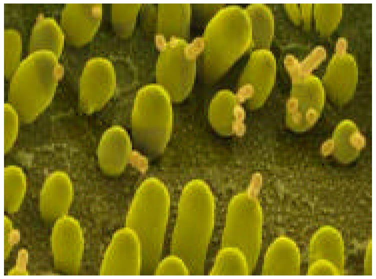
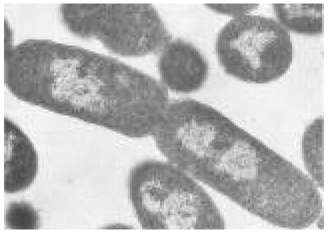
Comparative Genomics

Liping Wei, Ph.D.
Nexus Genomics, Inc.

Completely Sequenced Genomes

- 4 eucaryotes + (draft) human
- 9 archaea
- 32 bacteria
- >600 viruses

What is Comparative Genomics?

The practice of **analyzing and comparing** the **genetic material** of **different species** for the purpose of studying evolution, the functions of genes (what they do and why), and inherited diseases.

“Perhaps the most important achievement of the Human Genome Project is that it has spawned sequencing of other genomes from all walks of life.” -- EV Koonin, 2000

Why Comparative Genomics?

- It tells us what are common and what are unique between different species at the genome level.
  - One application is to identify unique, crucial proteins in pathogens to use as targets for products that are both safe and effective.
- Genome comparison may be the surest and most reliable way to identify genes and predict their functions and interactions.
  - e.g., to distinguish orthologs from paralogs
- The functions of human genes and other DNA regions can be revealed by studying their counterparts in lower organisms.

Three Major Research Directions

1. Genome comparison for the purpose of understanding the similarity and difference between the genomes
2. Genome comparison for the purpose of predicting gene function, exons, etc., of a new genome, and ultimately, the study of evolution.
3. Development of efficient algorithms for comparing large, genome-scale sequences.

Think Genome Scale…

DNA → Protein → Function → Organism → Population

- From single gene to whole genome, with increase in both size and complexity.
- From traditional homology-based approaches to new nonhomology-based approaches
- Promising technologies yet very new. Always question the assumptions.

Outline of Lecture

1. Comparison of complete genome sequences

   of two strains of *H. Pylori* to study strain-specific genetic diversity.

   "What are the features to be compared?"

2. Prediction of protein interaction maps

   for complete genomes based on gene fusion events.

   "What can we do with genome comparison?"

3. Relatively fast alignment of whole genome sequences

   using suffix tree

   "How to align genome-scale sequences?"

Outline of Lecture

1. **Comparison of complete genome sequences**

   of two strains of *H. Pylori* to study strain-specific genetic diversity.

   "What are the features to be compared?"

2. Prediction of protein interaction maps

   for complete genomes based on gene fusion events.

   "What can we do with genome comparison?"

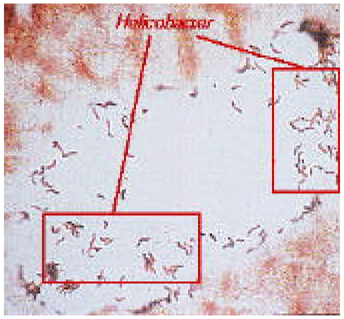
3. Relatively fast alignment of whole genome sequences

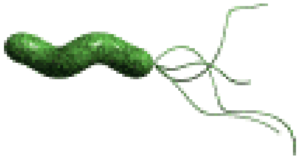
   using suffix tree.

   "How to align genome-scale sequences?"

Helicobacter pylori

- Colonizes the human gastric mucosa
- Induces chronic gastric inflammation which can progress to ulcer, gastric cancer or mucosal-associated lymphoma
- Affects 30-40% of population in US, 60-80% in Asia





H. Pylori can cause different diseases or even be beneficial to the infected host.

What causes the difference? Strain-specific genetic diversity, or host diversity?

RA Alm, et. al., 1999, Nature, 397: 176-80

- compare the genomes of two *H. Pylori* strains
- Strain J99 and strain 26695, two independent isolates.

What to compare?   1. Statistics of the genome

- **Size of the genome**: total number of base pairs
- **Overall (G+C) content**: percentage of (G+C)
- **Regions of different GC content**: (G+C) content in sliding windows
  - Are they the **corresponding regions** in both genomes?

Genome features	***H. Pylori* 26695**	***H.Pylori* J99**
Size (base pairs)	1,667,867	1,643,831
(G+C) content %	39	39
Regions of different (G+C) content	8	9*

* Four of the regions match those in 26695.

Nature, Vol 397: 176-80, 1999

What to Compare? 2. Predicted ORFs

How to identify genes in a genome?

- Accurate identification of genes in procaryotes and unicellular eucaryotes can be achieved by
  - homology to known genes in other species - ~80% of genes
  - Statistical methods: GenMark, Glimmer
- Accuracy is much poorer for multicellular eucaryotes, especially human.
  - Order-of-magnitude more difficult because of
    - Large and complex intron regions
    - Alternative splicing
  - Statistical methods: GenScan, Genie
  - Statistical analysis + homology: PROCRUSTES
  - + mRNA sequences and homology with other close genomes
- Manual adjustment is often required as the last step.

What to Compare? 2. Predicted ORFs

- **Total number** of predicted Open Reading Frame
- **Percentage of the Genome** (coding)
- **Average length**

- predicted genes with homology and **assigned function**
- predicted genes with **homology but no function**
- ***H. Pylori* specific** genes
- **Strain-specific** genes
- **Location of strain-specific** genes

What to Compare? 2. Predicted ORFs

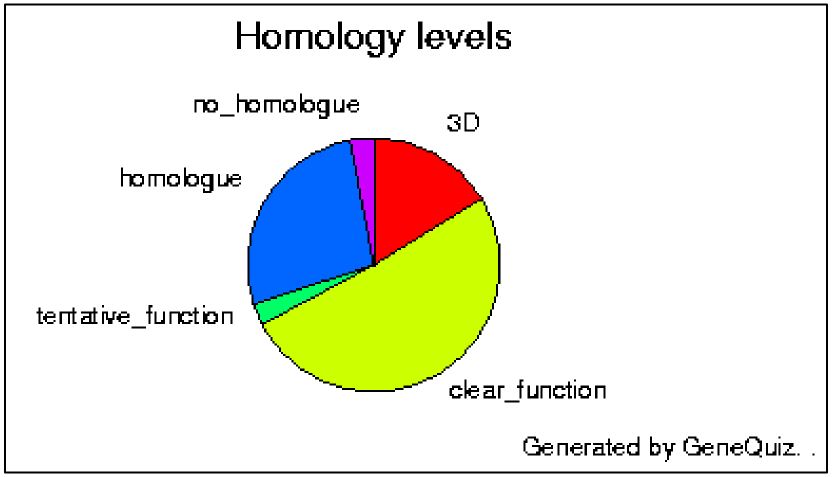
ORFs	***H. Pylori* 26695**	***H.Pylori* J99**
Total	1590	1495
Percentage of Genome (Coding)	91.0	90.8
Average length	954	998
Functionally classified	875	895
Conserved with no function	275	290
H. Pylori specific	345	367
Strain-specific genes *	117	89

* Half of the strain-specific genes are clustered in a plasticity zone with different (G+C) content, suggestive of horizontal DNA transfer.

Nature, Vol 397: 176-80, 1999

J99 Genome: Updated Function Assignment

Sequences/Assignment	Numbers	Percentage
sequences with a 3D homolog	242	16
function assigned by clear homology	1001	67
function assigned by tentative homology	41	2
homologue found but no function assigned	402	26
no homologue found	45	3
Total	1489	100



What to Compare? 3. Paralogues and Orthologues

- Paralogous families
- DNA-sequence differences between orthologues
- Protein-sequence differences between orthologues

- In J99, 337 genes are members of 113 paralogous families
- DNA-sequence differences between orthologues are mainly found in the third position of coding triplets
  - 8 genes with >98% nucleotide identity
  - 310 proteins with >98% amino-acid identity.

Nature, Vol 397: 176-80, 1999

What to Compare?
4. Genomic Organization and Gene Order

- Duplication
- Inversion and Translocation
- Gene order: conservation of immediate neighbors

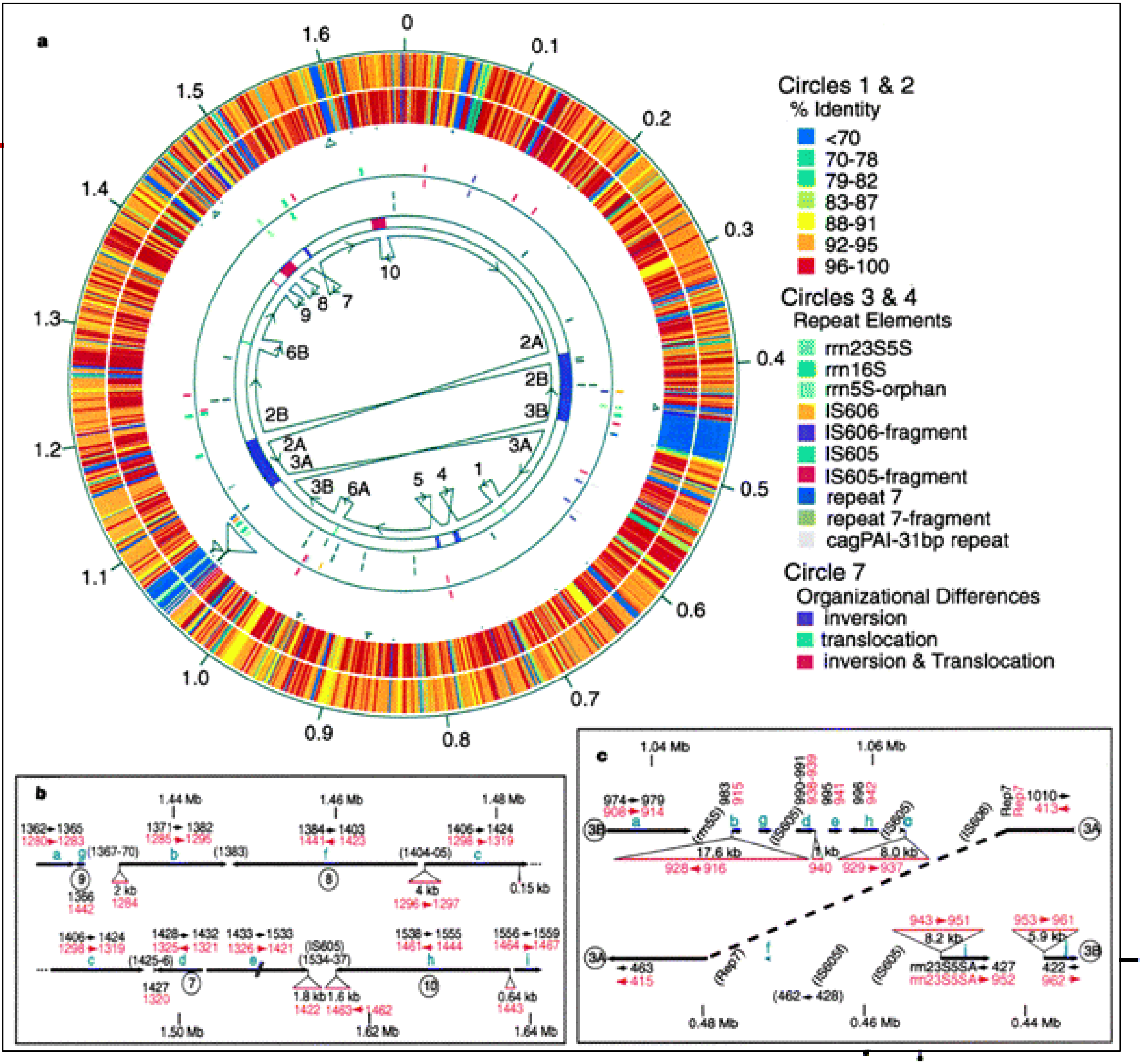
Alignment, alignment, alignment.

What to Compare?
4. Genomic Organization and Gene Order

- Three single-copy genes in 26695 have complete or partial duplications in J99
- 10 regions of inversion and/or translocation
- Gene Order:
    - 84.7% of the genes in J99 have the same neighbor on each side in both genomes.
    - 13.5% are flanked by strain-specific genes on one or both sides
    - Only 1.8% have a different neighbor on one side because of organizational differences

Nature, Vol 397: 176-80, 1999

Nature, Vol 397:
176-80, 1999



Outline of Lecture

1. Comparison of complete genome sequences

   of two strains of H. Pylori to study strain-specific genetic diversity.

   “What are the features to be compared?”

2. **Prediction of protein interaction maps**

   for complete genomes based on gene fusion events.

   “What can we do with genome comparison?”

3. Relatively fast alignment of whole genome sequences

   using suffix tree.

   “How to align genome-scale sequences?”

Detecting Protein Interaction

- Lives of biological cells are controlled by interacting proteins in metabolic and signaling pathways.
- Protein interactions are traditionally detected using experimental methods
  - Biochemistry: co-immunoprecipitation or crosslinking
  - Molecular biology: two-hybrid system or phage display
  - Genetics: unlinked noncomplementing mutant detection
- Computational method based on:
  - Subunit interfaces in protein structure databases
  - Gene order
  - Phylogenetic profile
  - Gene fusion

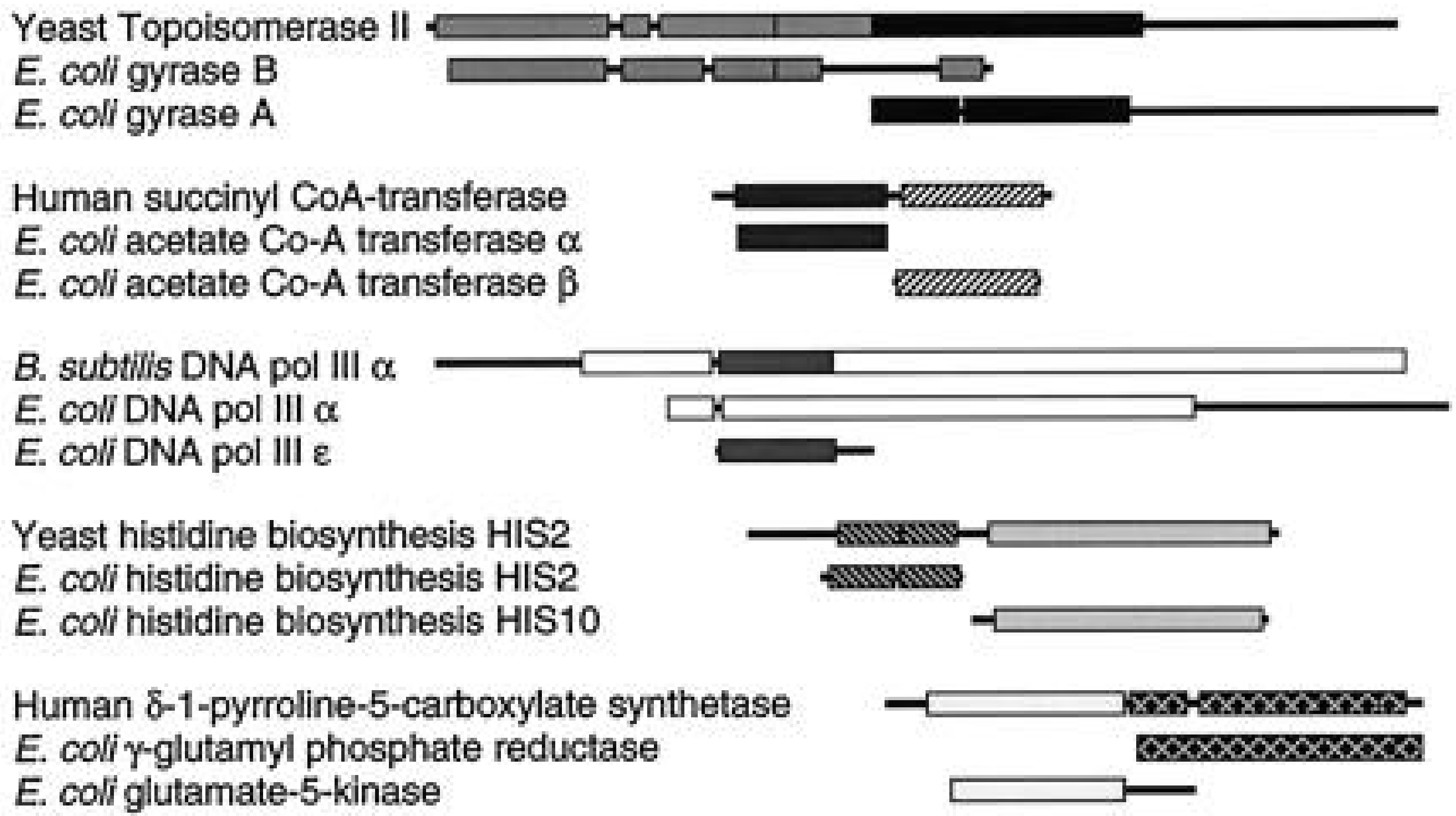
Predicting Protein Interaction Based on Gene Fusion

Definitions:

- **Gene fusion event**: certain protein families in a given species consist of **fused domains** that usually correspond to two or more single, full-length proteins in other species
- **Interaction** here is defined as either direct physical interaction or an indirect functional association (e.g., involvement in the same biochemical pathway or similar gene regulation)

Assumption: If a composite protein is uniquely similar to two component proteins in another species, the component proteins are most likely to interact.

Examples of Gene Fusion Events


Yeast Topoisomerase II
E. coli gyrase B
E. coli gyrase A
Human succinyl CoA-transferase
E. coli acetate Co-A transferase α
E. coli acetate Co-A transferase β
B. subtilis DNA pol III α
E. coli DNA pol III α
E. coli DNA pol III ε
Yeast histidine biosynthesis HIS2
E. coli histidine biosynthesis HIS2
E. coli histidine biosynthesis HIS10
Human δ-1-pyrroline-5-carboxylate synthetase
E. coli γ-glutamyl phosphate reductase
E. coli glutamate-5-kinase

Science, Vol 285: 751-3, 1999

Method

- Input: translation of all ORFs in complete genomes. One genome as query, and the other as references.
- Procedure:

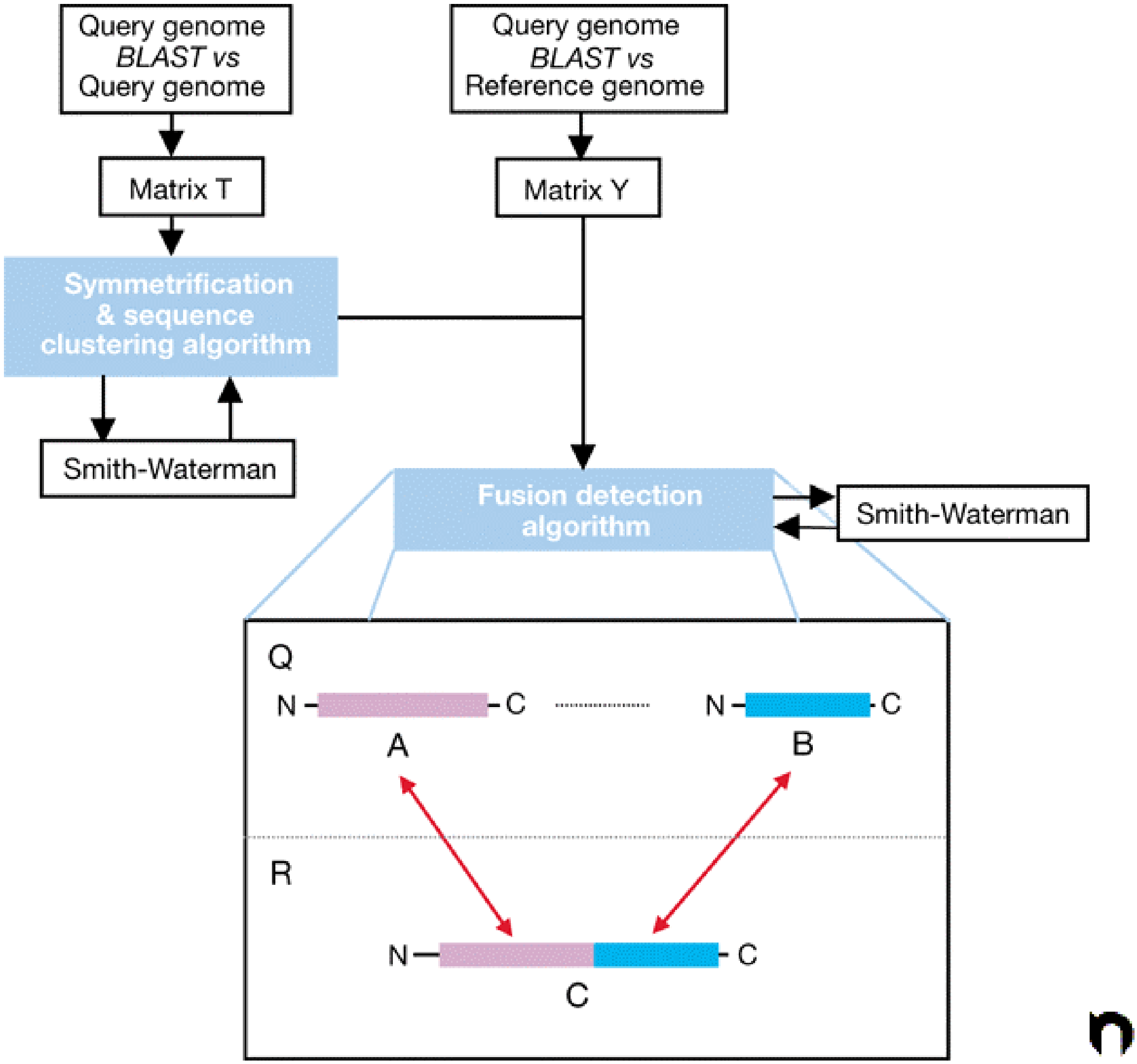
1. The query set is compared against itself using BLASTP; Pairwise sequence similarities are recorded in a binary matrix T.
   - mask compositionally biased regions (CAST)
   - Use Smith-Waterman to symmetrify the matrix
2. The query set is compared against a reference set using BLASTP; Pairwise sequence similarities are recorded in binary matrix Y

Method (Continued)

- Procedure (continued):

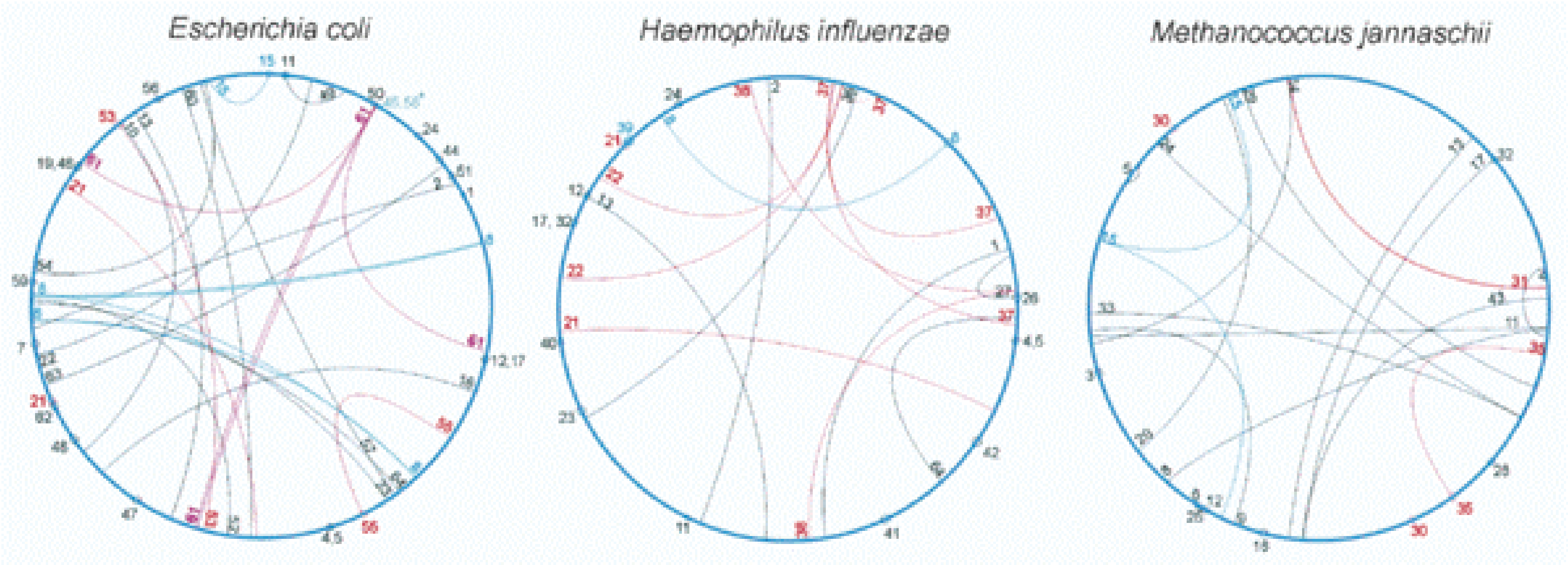
  3. For each entry C in reference set, collect pair (A,B) from the query set where both A and B are similar to C.
     - Look up (A,B) in Matrix T.
     - If (A, B) is null in T, run Smith-Waterman to confirm dissimilarity
     - If dissimilar, collect (A,B) as candidates for a fusion event

Method Flowchart



Nature, Vol 402: 86-90, 1999

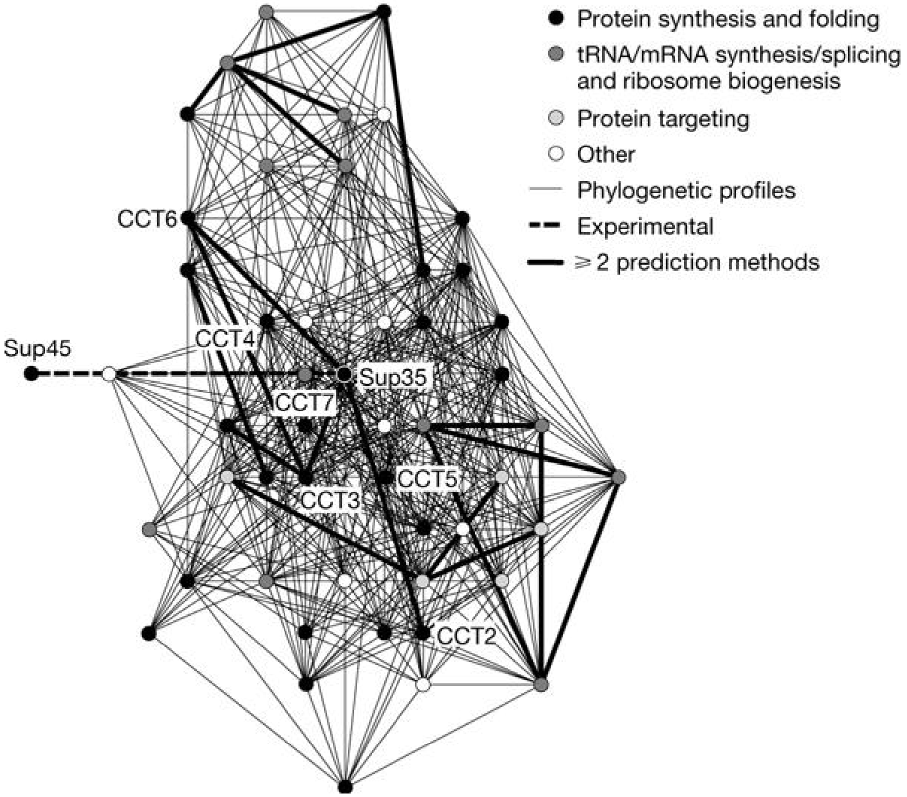
Predicted Protein Interaction Maps



- Four reference genomes: the above three + yeast genome
- Number of predicted interactions: 39, 24, and 25, respectively

Nature, Vol 402: 86-90, 1999

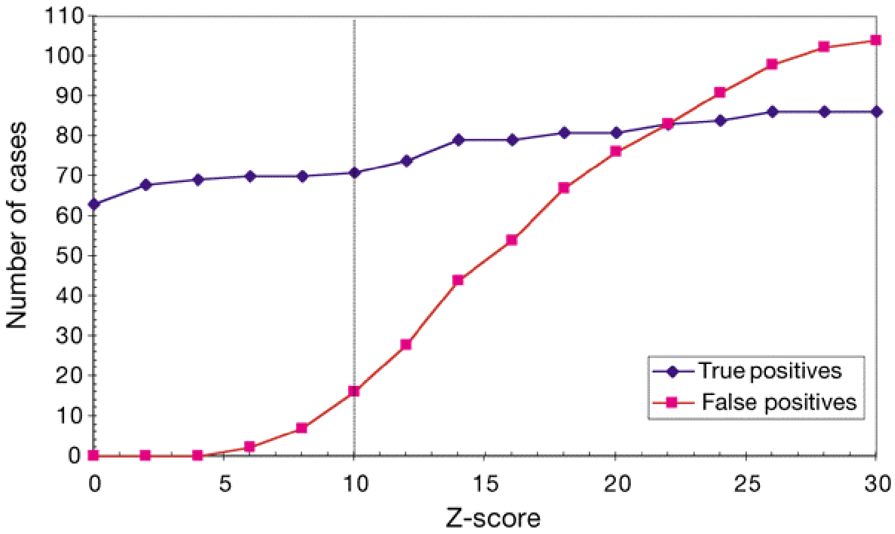
Predicted Protein Interaction Map for Yeast



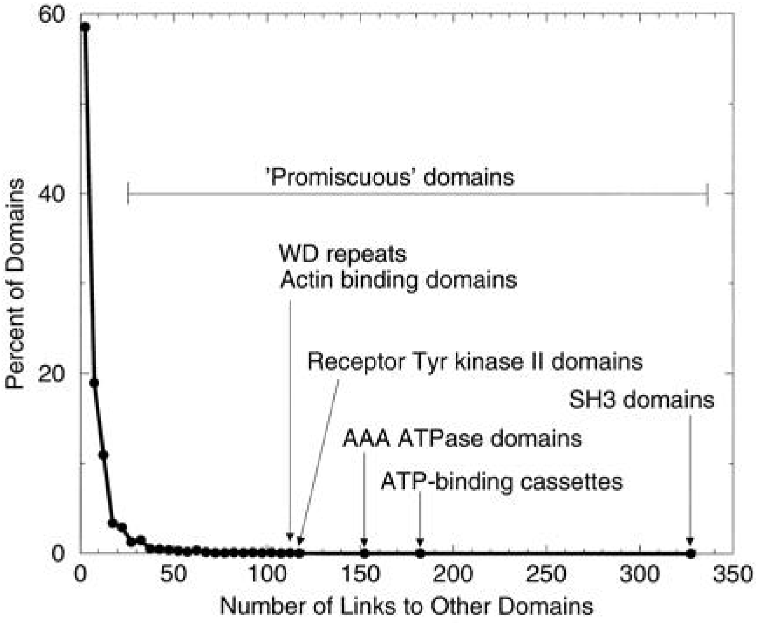
- 20 reference genomes
- Number of predicted interaction: 45,502

Nature, Vol 402: 83-6, 1999

Reduce False Positives



Nature, Vol 402: 86-90, 1999



Science, Vol 285: 751-3, 1999

Outline of Lecture

1. Comparison of complete genome sequences

   of two strains of H. Pylori to study strain-specific genetic diversity.

   "What are the features to be compared?"

2. Prediction of protein interaction maps

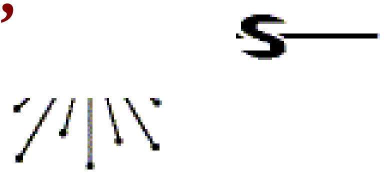
   for complete genomes based on gene fusion events.

   "What can we do with genome comparison?"

3. **Relatively fast alignment of whole genome sequences**

   using suffix tree.

   "How to align genome-scale sequences?"



Sequence Alignment—Genome Scale

- Single genes: thousands to tens of thousands of bases
- Single proteins: hundreds to thousands of residues
- Complete genomes:

Species	**Total number of base pairs in genome**
H. pylori	1.6 million
E. Coli	4.8 million
Baker’s yeast	12 million
C. elegans	97 million
Drosophila	137 million
Human	3.1 billion
Mouse	3.1 billion

Challenges

- Large size of the DNA sequences to be aligned
  - Memory
  - Speed
- Occurrence of both short and long insertions and deletions
- Large-scale changes such as tandem repeats and large-scale reversals
- High degree of divergence in the third position of codons

A Suffix tree-based Method

- AL Delcher, et. al., 1999, Nucleic Acids Research
- Designed for fast alignment of large, closely-related sequences
  - Assumption: there is a mapping between large subsequences of the two inputs
- Aligned two ~4Mb genome sequences of two tuberculosis strains in less than 1 min of computation time

Three Steps

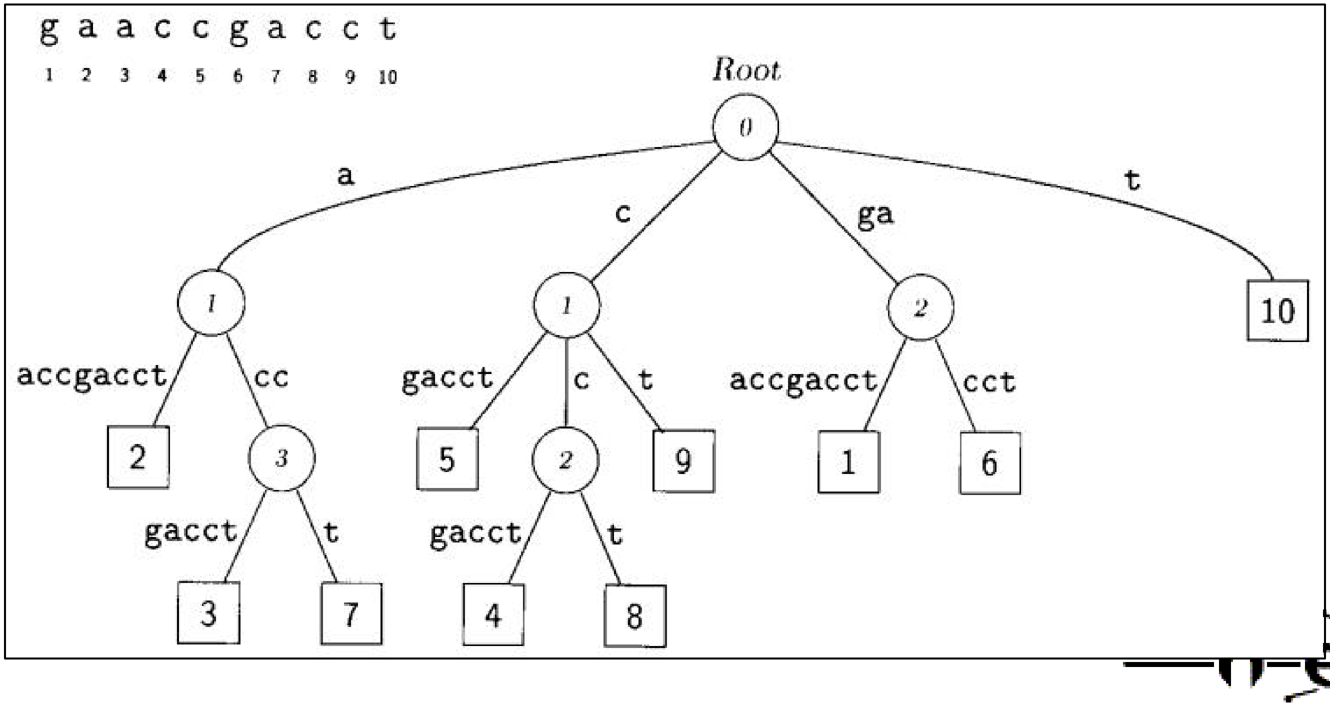
1. Identify all Maximal Unique Matches (MUMs)
2. Extract the longest set of matches that occur in the same order in both genomes
3. Close the local gaps by identifying inserts, repeats, tandem repeats, small mutated regions, and SNPs

Step 1: Identify all MUMs

- A **Maximal Unique Match** is a subsequence that occurs exactly once in Genome A and once in Genome B, and is not contained in any longer such sequence.
- A naïve algorithm is O(n^3), where n is the sum of the length of Genome A and B.
- Use the suffix tree data structure for efficiency → O(n) for both run time and space
  - A generous upper bound for space: 37 bytes per base.
  - << 8 Gb of memory for comparison of two 100 Mb sequences

Suffix Tree

- A suffix tree is a compact representation that stores all possible suffixes of an input sequence S.
- A suffix is a subsequence that begins at any position in the sequence and extends to the end of the sequence.



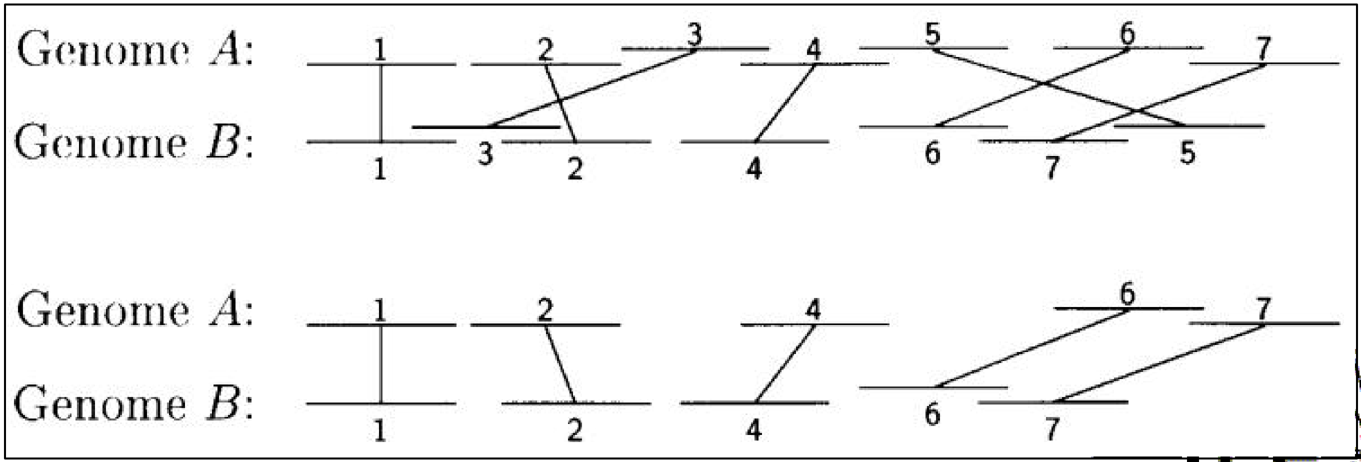
Nucleic Acid Research, 27(11):2369-76, 1999

Suffix Tree

- Concatenate the two genomes into one sequence separated by a dummy character.
- Use McCreight's algorithm to build suffix tree in linear time.
  - Clever use of sets of pointers
- Label each leaf node to indicate which suffix it represents in which genome.
- Identify all Maximal Unique Matches in one scan
  - Every unique matching sequence is represented by an internal node with exactly two child leaf nodes, one from each genome.
  - Unique matches that are maximal can be identified by mismatches at their ends
- **Identify MUMs on both DNA strands**

Step 2: Sorting the MUMs

- Set length of the shortest MUM.
  - e.g., 50 for highly similar genomes, 20 for similar ones
- Sort the MUMs according to their position in Genome A
- Use a variation of the Longest Increasing Subsequence algorithm.
- Run time O(KlogK), where K= number of MUMs



Nucleic Acid Research, 27(11):2369-76, 1999

Step 3: Closing the Gaps – four classes

1. Repeats
   - In their study, most repeats were tandem repeats. All their tandem repeats were adjacent to unique sequence
   - Can be identified by MUMs overlapping each other
2. SNPs
   - Simple case: gap of one base between MUMs.
   - SNP adjacent to repeat sequences: use repeat processing
3. Insertions
   - Simple insertion: large gap in alignment in one genome but not the other
   - Transposition: appear in MUM alignment out of sequence.
4. Variable/Polymorphic regions:
   - Appear as gap in MUM alignment
   - If short, use Smith-Waterman dynamic programming
   - If long, run MUM detection with reduced minimum length.

Computation time vs. size and similarity

	Length	Sequence similarity	Step 1 (# sec)	Step 2 (# sec)	Step 3 (# sec)
M. Tuberculosis H37Rv Vs. *M. Tuberculosis* CDC1551	4Mb 4Mb	99% identical	5	45	5
M. Genitalium *vs.* *M. pneumoniae*	580Kb 816Kb	20Kb in MUMs of >15b; < 50%id in gap regions	6.5	0.02	116
Subsequences of Human chromosome12p13 vs. Mouse chromosome 6	223kb 228kb	14kb in MUMs of >15b; Large gaps	1.6	29	?

Pro and Con

- Pros:
  - very fast for alignment of genomes of different strains of the same species or genomes of similar species
  - Can handle long insertions and deletions
  - Can detect reverses, SNPs, repeats, and tandem repeats

- Con:
  - speed suffer significantly for less similar sequences
    - Minimum MUM length needs to be set lower
    - Many more runs of Smith-Waterman in Step 3

Another Genome-Scale Alignment Method: WABA

- WJ Kent, AM Zahler, 2000, Genome Research
- Three passes
  - Identify homologous regions
  - Align in detail overlapping 2000x5000 base regions
  - Join the overlapping alignments
- Aligned 8 million bases of *Caenorhaditis briggsae* against the entire 97 million bases of *Caenorhaditis elegans* genome.
  - Overall similarity: 59% sequence identity.
- Run time on a Pentium III 450 mHz,
  - First pass: 20 hrs. O(MN)
  - Second pass: 11 days. O(min{M, N})
  - Third pass: 15 min. O(min{M, N})

Other Research Areas in Comparative Genomics

- Using genome comparison for exon prediction and regulatory region prediction
- Building phylogenetic tree based on genome comparison
- Visualization of genome alignment
- And more…

Summary

- Comparative genomics is a very powerful tool to study organism diversity, evolution, gene function, and etc.
- Think genome scale.
- Because they are new, many techniques need to be further validated. Be critical--always question the assumptions.

Liping Wei: wei@nexusgenomics.com