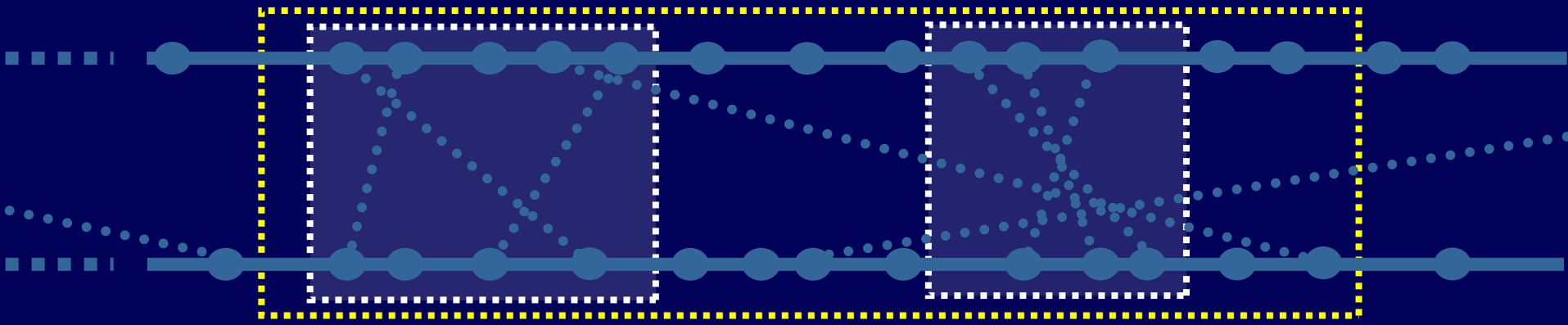
The Incompatible Desiderata of Gene Cluster Properties

Rose Hoberman
Carnegie Mellon University

joint work with Dannie Durand

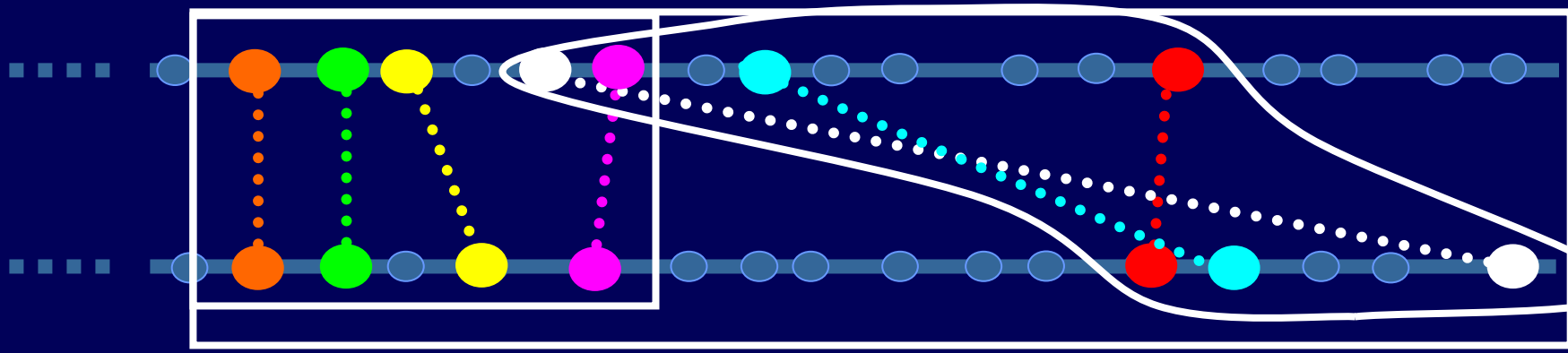
How to detect segmental homology?



- Intuitive notions of what gene clusters look like
  - Enriched for homologous gene pairs
  - Neither gene content nor order is perfectly preserved

How can we define a gene cluster formally?

Definitions will be application-dependent



- If the goal is to estimate the number of inversions, then gene order should be preserved
- If the goal is to find duplicated segments, allow some disorder

Gene Clusters Definitions

Large-Scale Duplications

Vandepoele et al 02

McLysaght et al 02

Hampson et al 03

Panopoulou et al 03

Guyot & Keller, 04

Kellis et al, 04

...

Genome rearrangements

Bourque et al, 05

Pevzner & Tesler 03

Coghlan and Wolfe 02

...

Functional Associations between Genes

Tamames 01

Wolf et al 01

Chen et al 04

Westover et al 05

...

Algorithmic and Statistical Communities

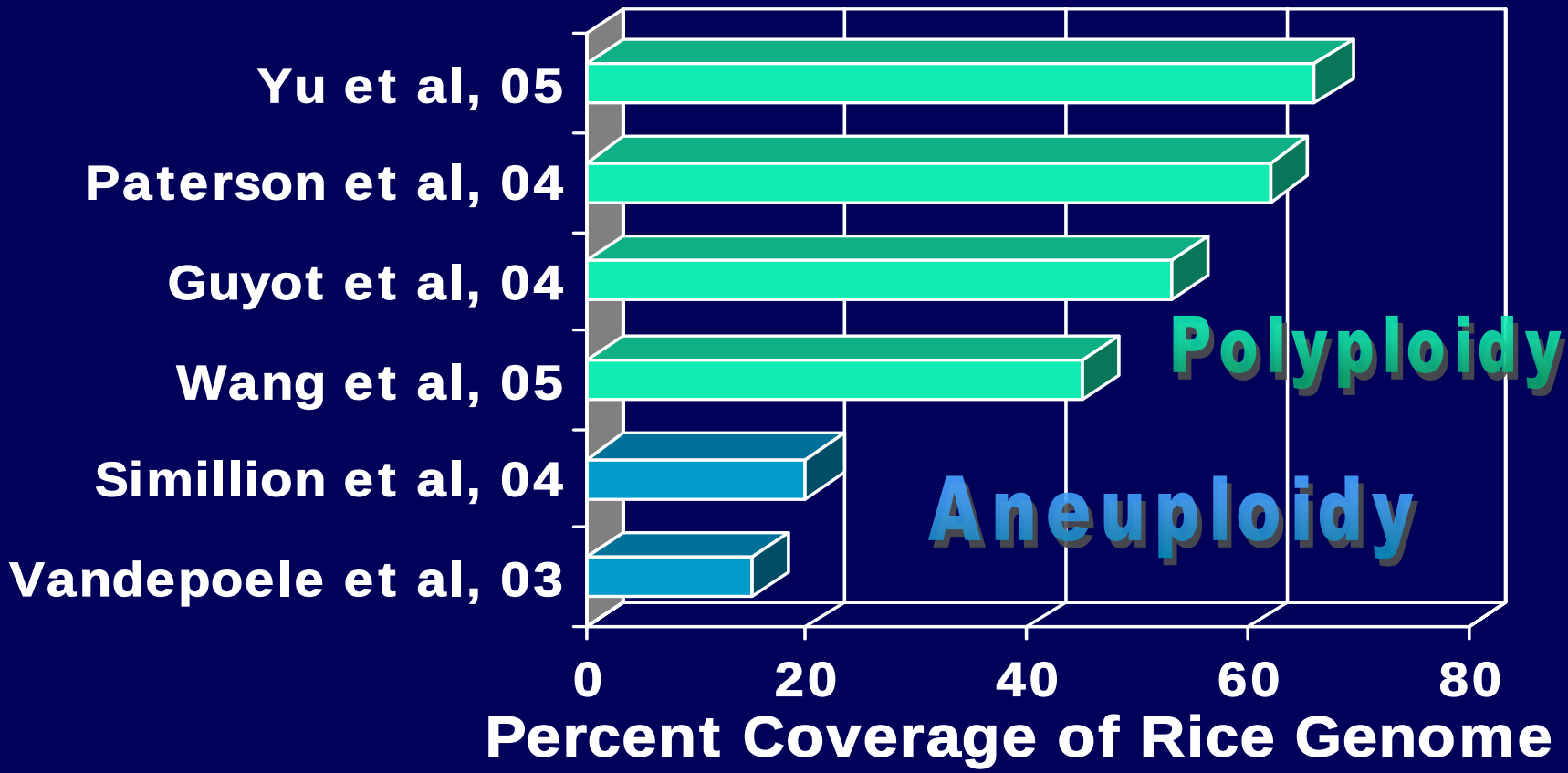
Bergeron et al 02

Calabrese et al 03

Heber & Stoye 01

...

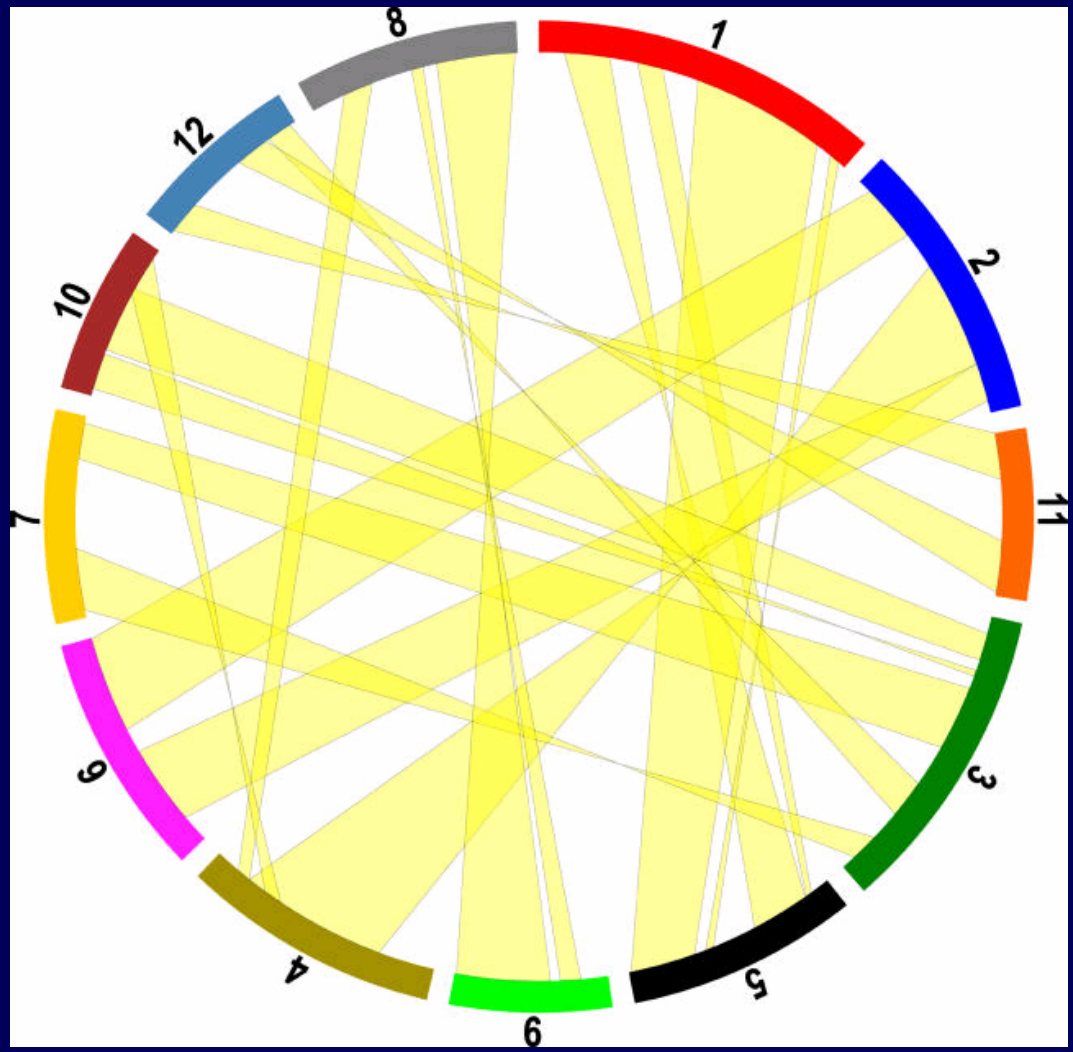
Groups find very different clusters when analyzing the same data



Cluster locations differ from study to study

⇩

Inference of duplication mechanism for individual genes varies greatly



The Genomes of *Oryza sativa*: A History of Duplications
Yu et al, PLoS Biology 2005

Goals:

Characterizing existing definitions

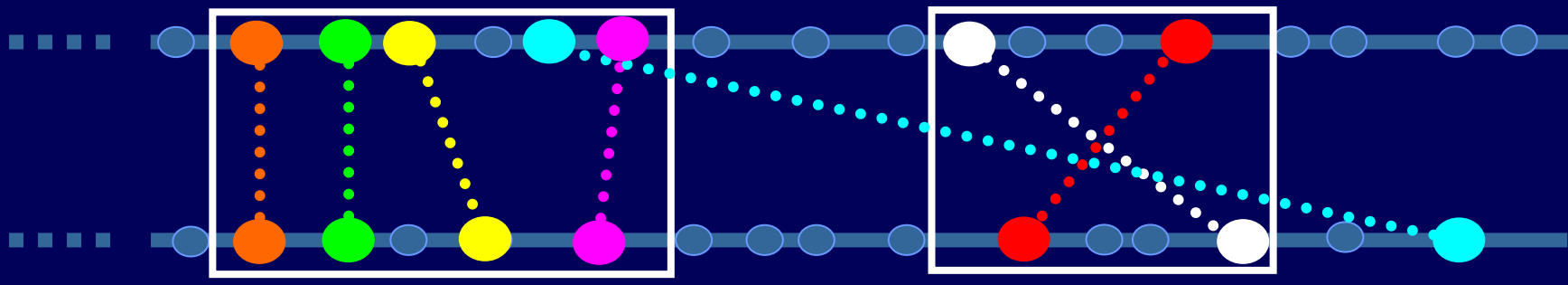
Formal properties form a basis for comparison

Gene cluster desiderata

Outline

- Introduction
- Brief overview of gene cluster identification
- Proposed properties for comparison
- Analysis of data: nested property

Detecting Homologous Chromosomal Segments (a marker-based approach)



1. Find homologous genes
2. Formally define a “gene cluster”
3. Devise an algorithm to identify clusters
4. Statistically verify that clusters indicate common ancestry

Cluster definitions in the literature

Descriptive:

- r-windows
- connected components (Pevzner & Tesler 03)
- common intervals (Uno and Tagiura 00)
- max-gap
- …

Require search algorithms

Constructive:

- LineUp (Hampson et al 03)
- CloseUp (Hampson et al 05)
- FISH (Calabrese et al 03)
- AdHoRe (Vandepoele et al 02)
- Gene teams (Bergeron et al 02)
- greedy max-gap (Hokamp 01)
- …

Harder to reason about formally

Cluster definitions in the literature

Descriptive:

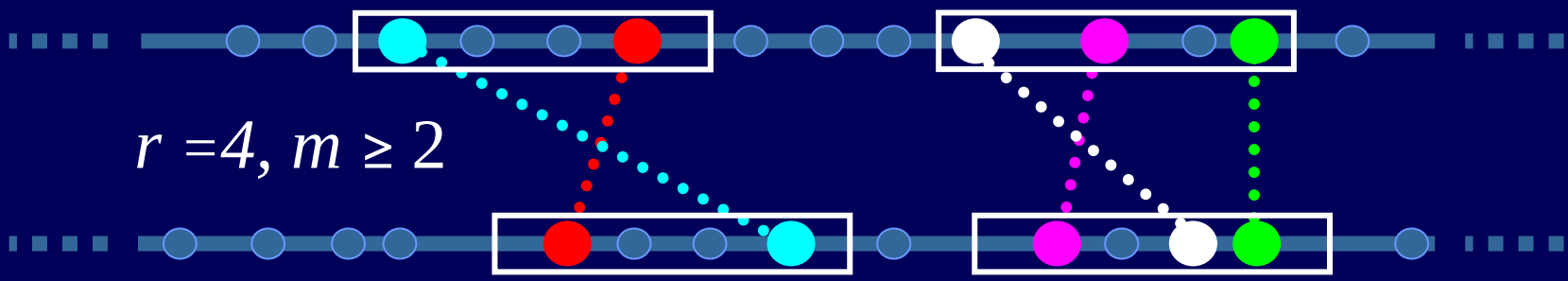
- r-windows
- connected components (Pevzner & Tesler 03)
- common intervals (Uno and Tagiura 00)
- max-gap
- …

Constructive:

- LineUp (Hampson et al 03)
- CloseUp (Hampson et al 05)
- FISH (Calabrese et al 03)
- AdHoRe (Vandepoele et al 02)
- Gene teams (Bergeron et al 02)
- greedy max-gap (Hokamp 01)
- …

I illustrate properties with a few definitions

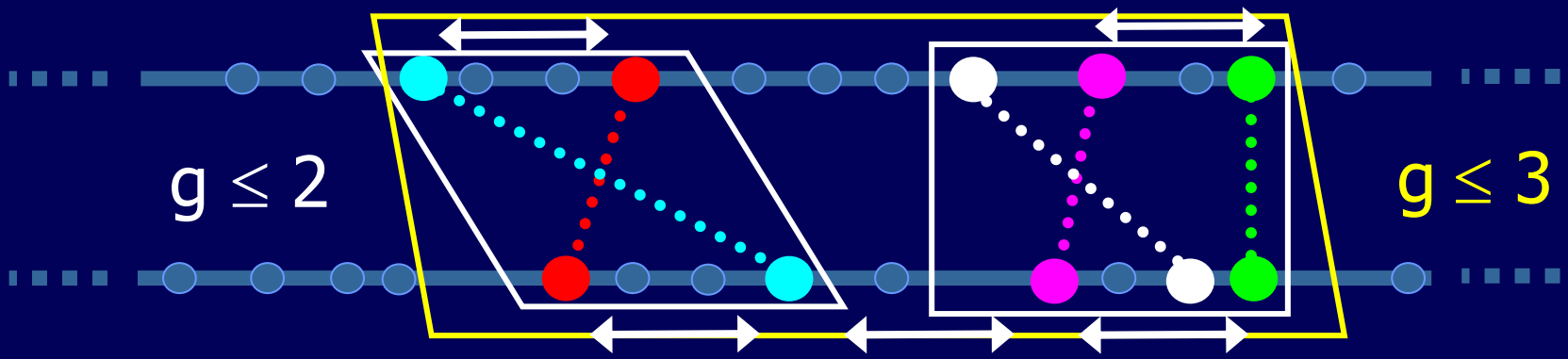
r-windows



- Two windows of size *r* that share at least *m* homologous gene pairs

(Calvacanti et al 03, Durand and Sankoff 03, Friedman & Hughes 01, Raghupathy and Durand 05)

max-gap cluster



A set of genes form a max-gap cluster if the gap between adjacent genes is never greater than *g* on either genome

Widely used definition in genomic studies

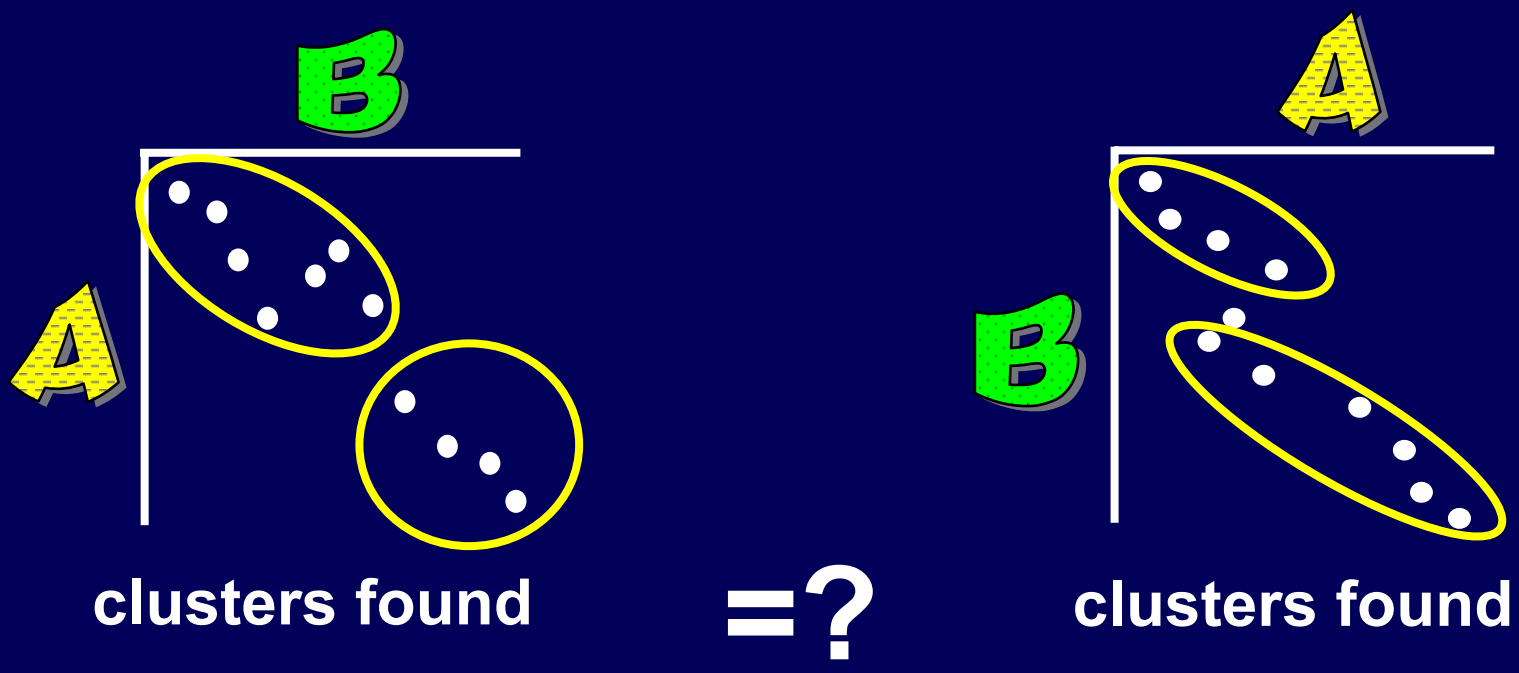
Outline

- Introduction
- Brief overview of existing approaches
- Proposed properties for comparison
- Analysis of data: nested property

Proposed Cluster Properties

- Symmetry
- Size
- Density
- Order
- Orientation
- Nestedness
- Disjointness
- Isolation
- Temporal Coherence

Symmetry



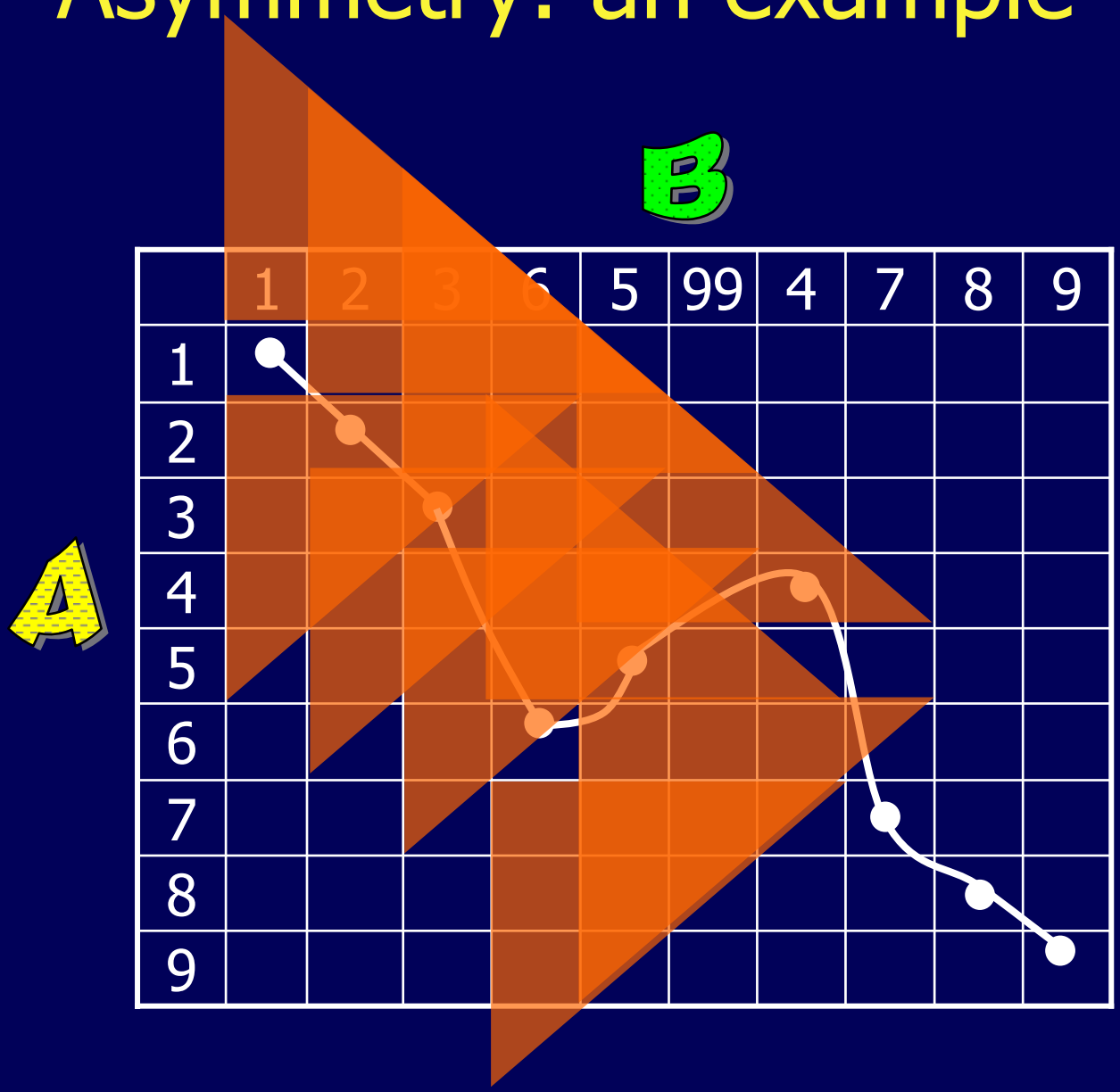
Many existing cluster algorithms are not symmetric with respect to chromosome

Asymmetry: an example

FISH (Calabrese et al, 2003)

- Constructive cluster definition: clusters correspond to paths through a dot-plot
- Publicly available software
- Statistical model

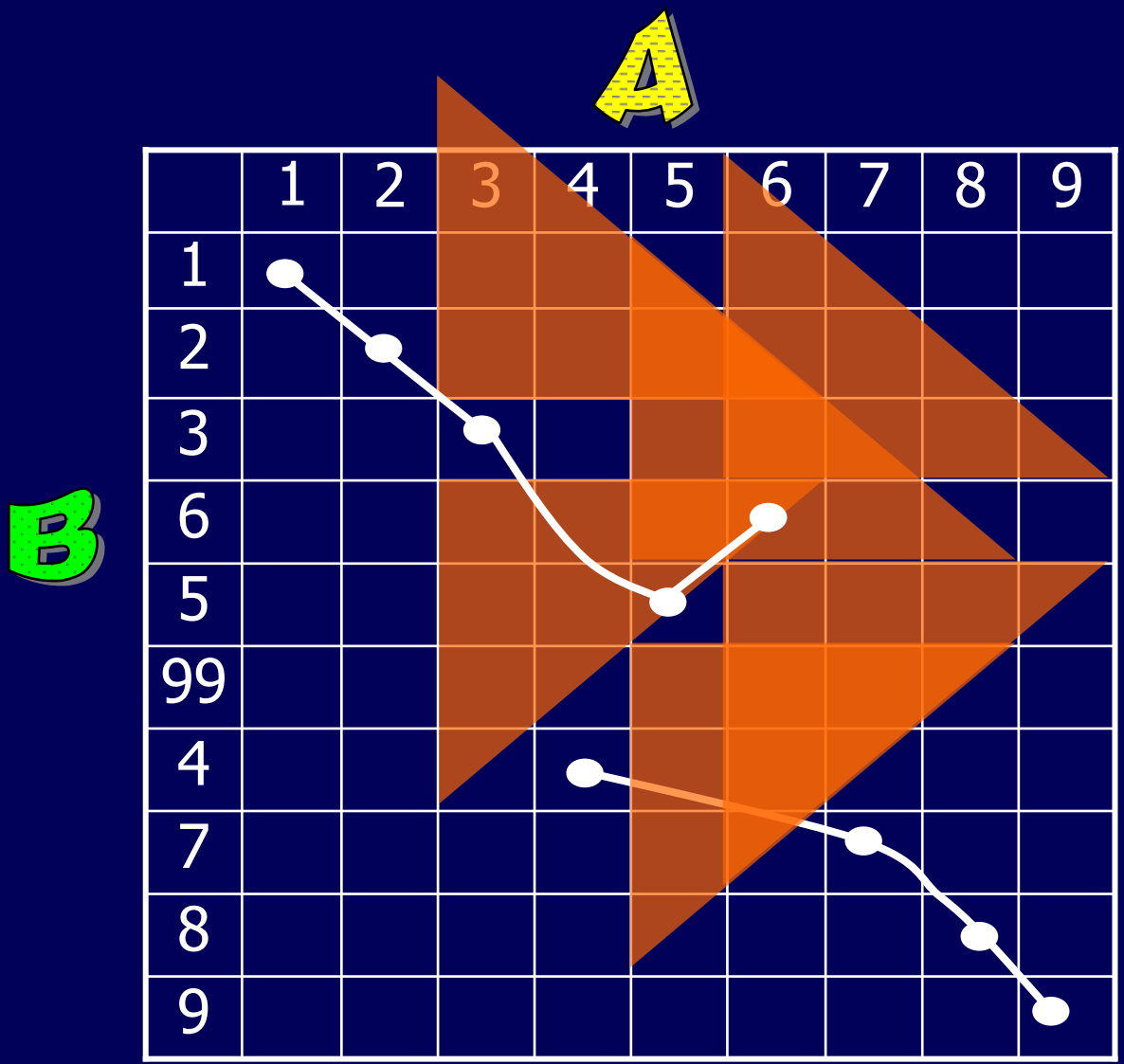
Asymmetry: an example



FISH

- Euclidian distance between gene pairs is constrained
- Paths in the dot-plot must always move to the right

Switching the axes yields different clusters



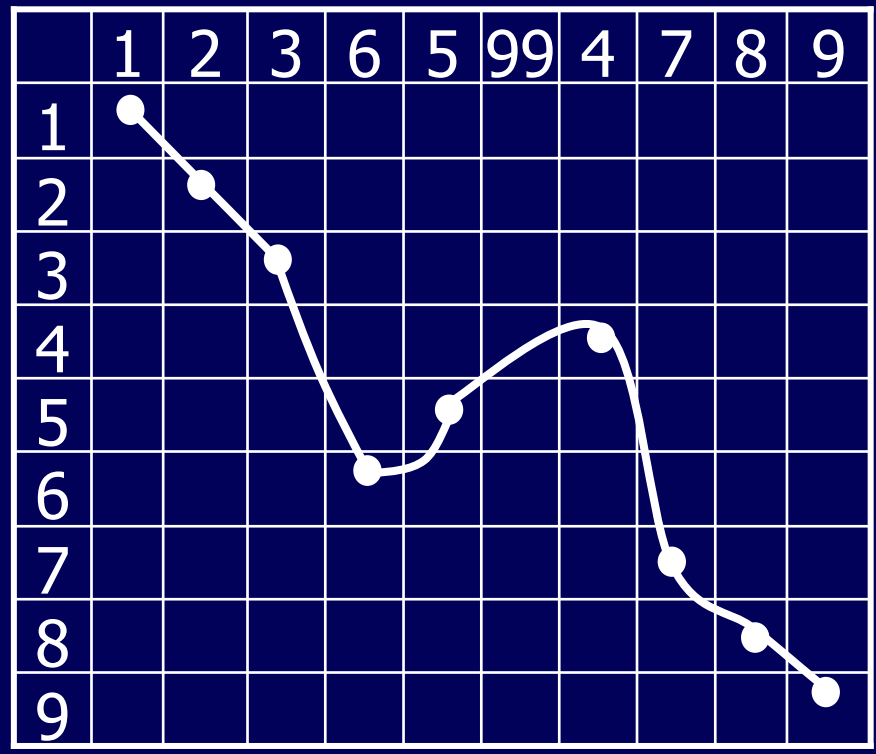
FISH

- Euclidian distance between markers is constrained
- Paths in the dot-plot must always move to the right

Ways to regain symmetry

1. Paths in the dot-plot must always move *down* and to the right
   - miss the inversion

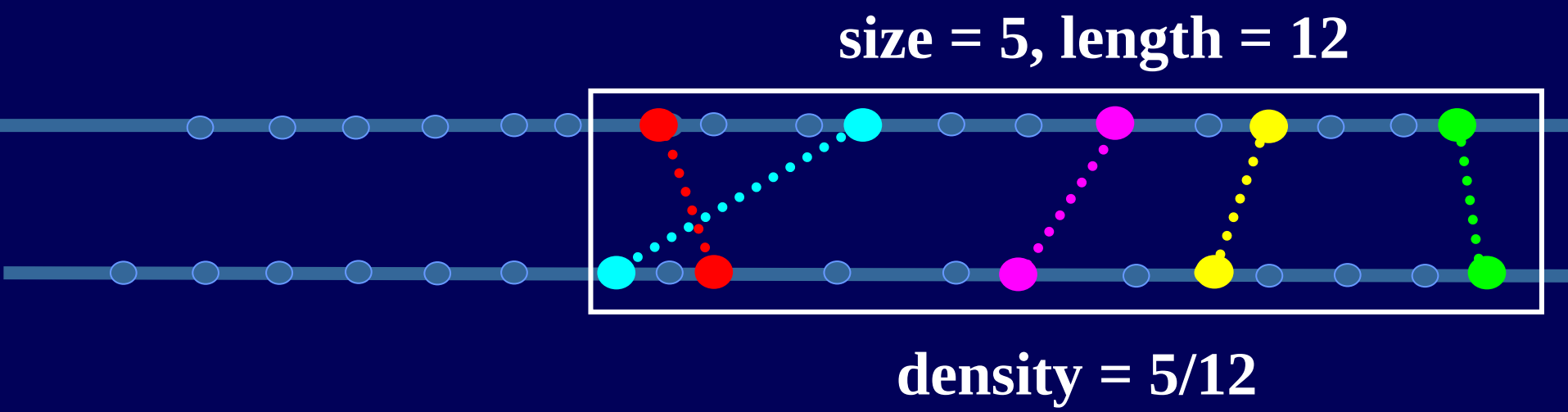
2. Paths can move in *any* direction
   - statistics becomes difficult



Regaining symmetry entails some tradeoffs

Proposed Cluster Properties

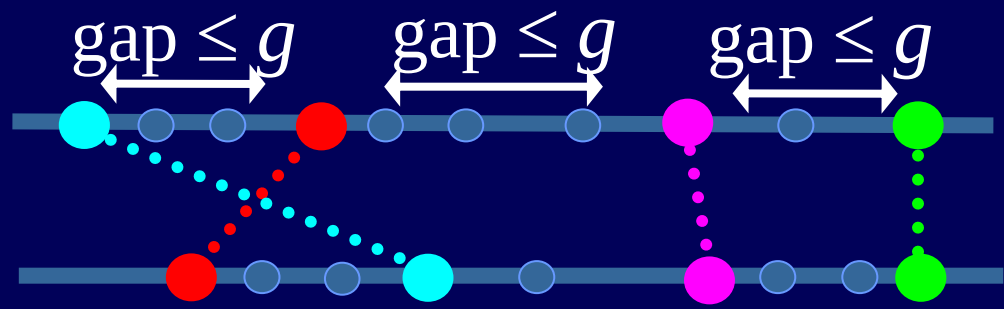
- Symmetry
- Size
- Density
- Order
- Orientation
- Nestedness
- Disjointness
- Isolation
- Temporal Coherence



Cluster Parameters

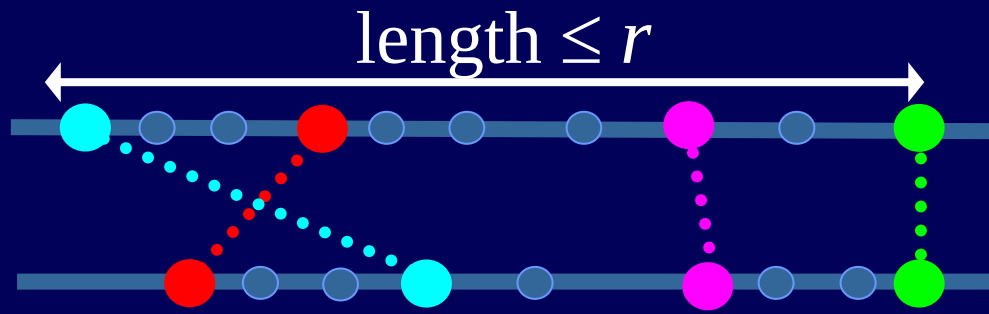
- size: number of homologous pairs in the cluster
- length: total number of genes in the cluster
- density: proportion of homologous pairs (size/length)

max-gap clusters



- cluster grows to its natural size
- cluster of size m may be of length m to $g(m-1)+m$
- maximal length grows as size grows

r-windows

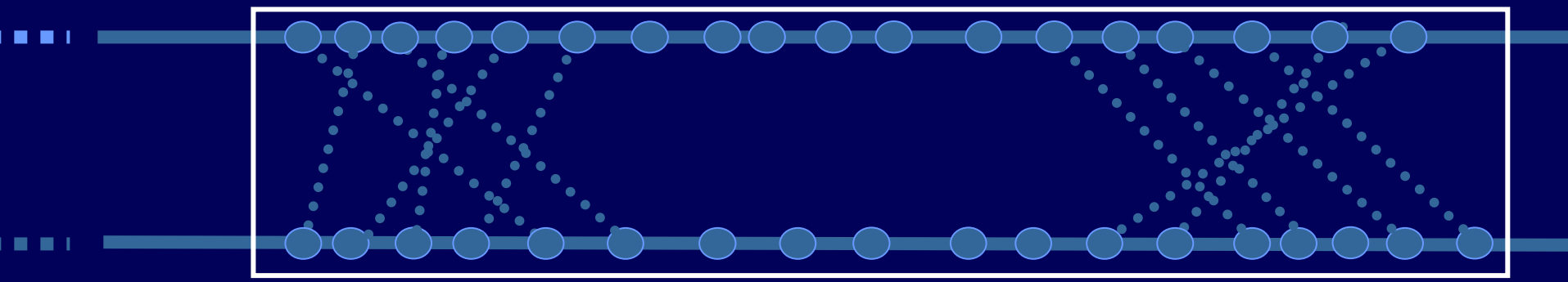


- cluster size is constrained
- cluster of size m may be of length m to r
- maximal length is fixed, regardless of cluster size

A tradeoff: local vs global density

- max-gap
  - constrains local density
  - only weakly constrains global density ($\geq 1/(g+1)$)
- r-window
  - constrains global density
  - only weakly constrains local density (maximum possible gap $\leq$ r-m)

Even when global density is high,



Density = 12/18

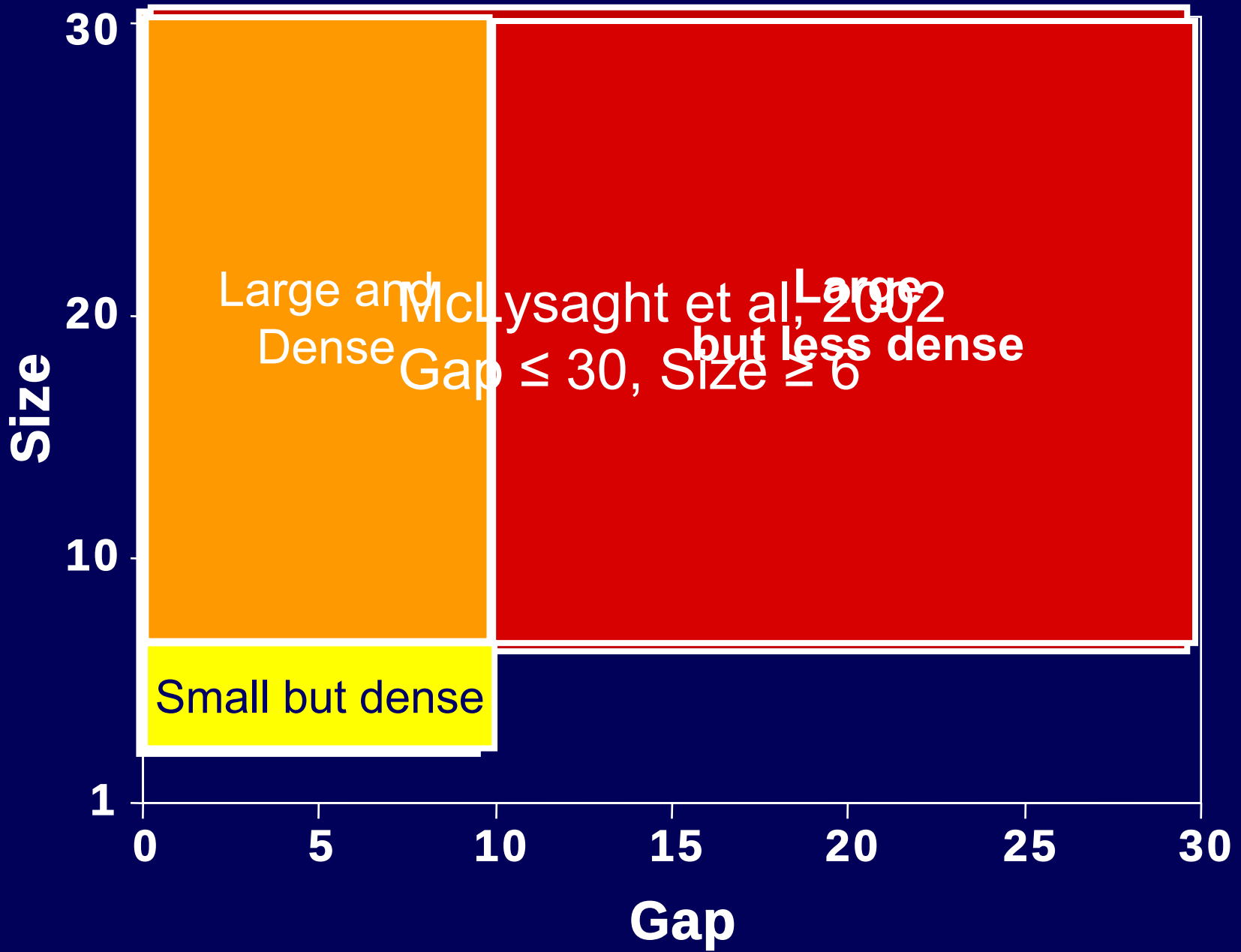
a region may not be locally dense

Size vs Density: An example

Application: all-against-all comparison of human chromosomes to find duplicated blocks

	Maximum Gap	Cluster Size	Post-Processing
McLysaght *et al*, 2002	constrained	test statistic	
Panopoulou *et al*, 2003	test statistic	constrained	merged nearby clusters

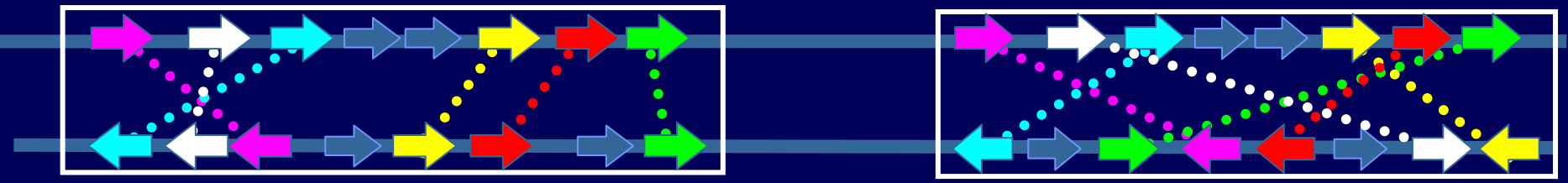
A Tradeoff in Parameter Space



Proposed Cluster Properties

- Symmetry
- Size
- Density
- Order
- Orientation
- Disjointntess
- Isolation
- Nestedness
- Temporal Coherence

Order and Orientation



density = 6/8

density = 6/8

- Local rearrangements will cause both gene order and orientation to diverge
  - Overly stringent order constraints could lead to false negatives
  - Partial conservation of order and orientation provide additional evidence of regional homology

Wide Variation in Order Constraints

- None (r-windows, max-gap, ...)
- Explicit constraints:
  - Limited number of order violations (Hampson et al, 03)
  - Near-diagonals in the dot-plot (Calabrese et al 03, ...)
  - Test statistic (Sankoff and Haque, 05)
- Implicit constraints: via the search algorithm (Hampson et al 05, ...)

Proposed Cluster Properties

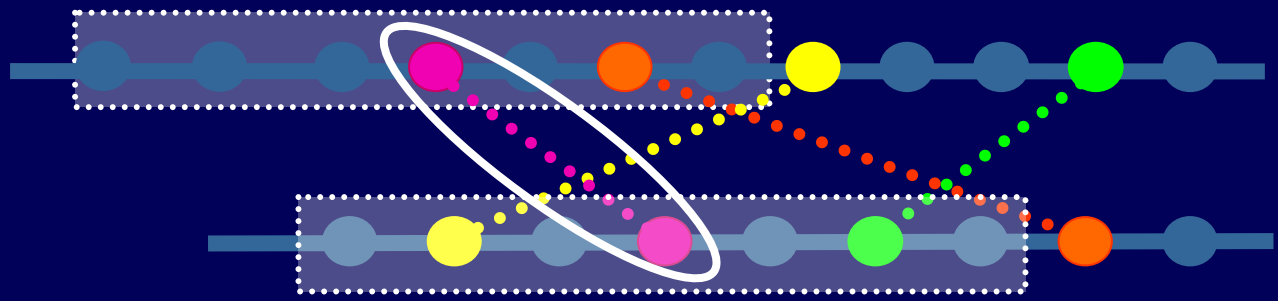
- Symmetry
- Size
- Density
- Disjointness
- Isolation
- Order
- Orientation
- Nestedness
- Temporal Coherence

Nestedness

- In particular, implicit ordering constraints are imposed by many greedy, agglomerative search algorithms
- Formally, such search algorithms will find only *nested* clusters

A cluster of size m is nested if
it contains sub-clusters of size m-1,...,1

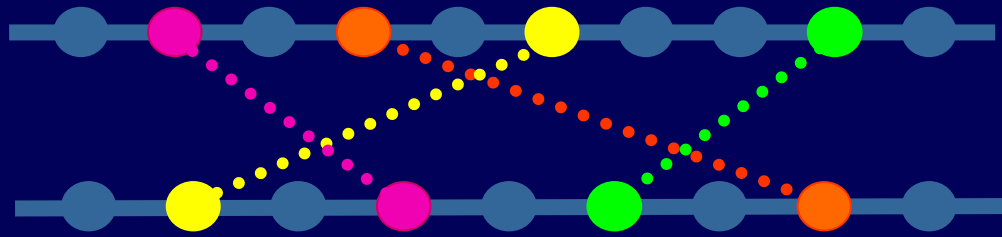
Greedy Algorithms Impose Order Constraints



$g = 2$

- A greedy, agglomerative algorithm
  - initializes a cluster as a single homologous pair
  - searches for a gene in proximity on both chromosomes
  - either extends the cluster and repeats, or terminates

Greediness: an example (Bergeron et al, 02)



$g = 2$

A max-gap cluster of size four

- No greedy, agglomerative algorithm will find this cluster
  - There is no max-gap cluster of size 2 (or 3)
- In other words, the cluster is not nested

Thus: different results when searching for max-gap clusters

- Greedy algorithms
  - agglomerative
  - find *nested* max-gap clusters

- Gene Teams algorithm (Bergeron *et al* 02; Beal *et al* 03,...)
  - divide-and-conquer
  - finds *all* max-gap clusters, nested or not

An example of a greedy search: CloseUp (Hampson et al, Bioinformatics, 2005)

- Software tool to find clusters
- Goal: statistical detection of chromosomal homology using *density alone*
- Method:
  - greedy search for nearby matches
  - terminates when density is low
  - randomization to statistically verify clusters

A comparative study
(Hampson et al, 05)

Is order information necessary
or even helpful for cluster detection?

- Empirical comparison:
  - CloseUp: "density alone", but *greedy*
  - LineUp and ADHoRe: density + order information
  - evaluated accuracy on synthetic data

A comparative study
(Hampson et al, 05)

Is order information necessary
or even helpful for cluster detection?

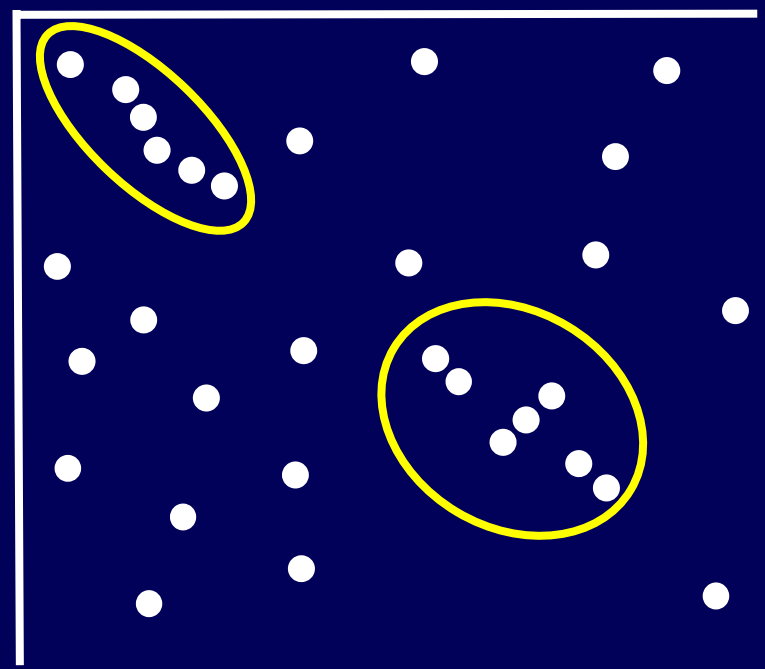
- Result: CloseUp had comparable performance
- Their conclusion: order is not particularly helpful
- My conclusion: results are actually inconclusive, since CloseUp implicitly constrains order

Proposed Cluster Properties

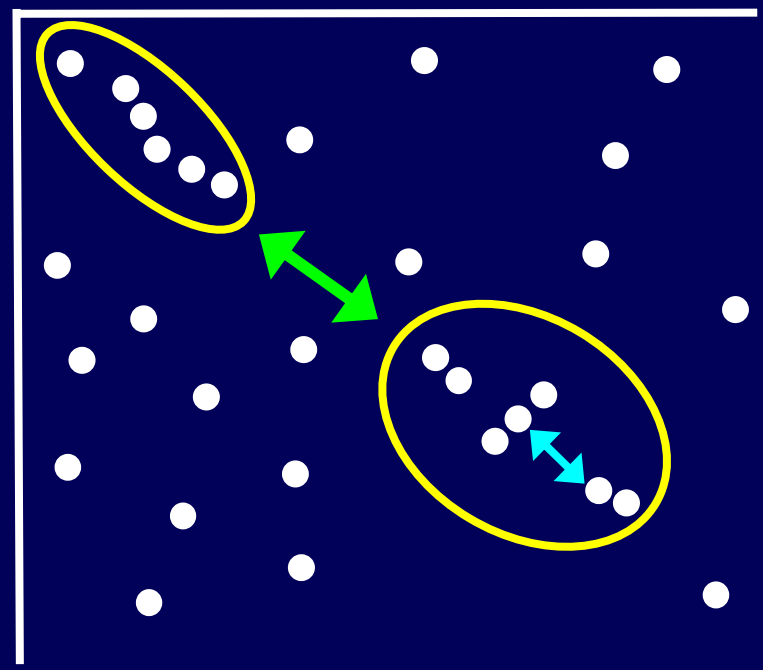
- Symmetry
- Size
- Density
- Order
- Orientation
- Nestedness
- Disjointness
- Isolation
- Temporal Coherence

Gene clusters: islands of homology in a sea of interlopers

How can we formally describe this intuitive notion?



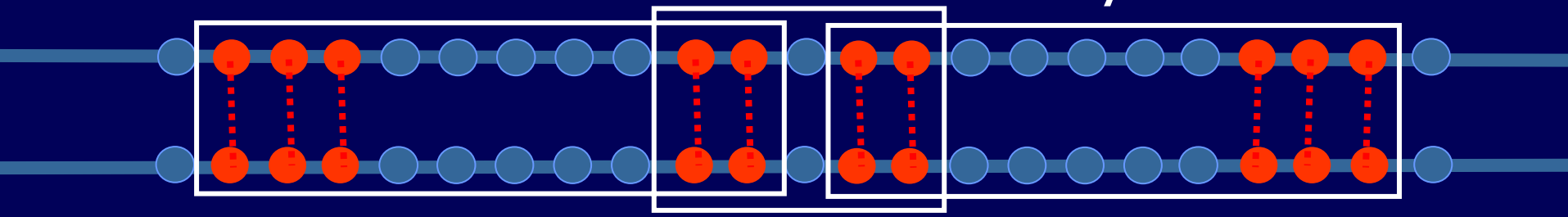
Islands of Homology



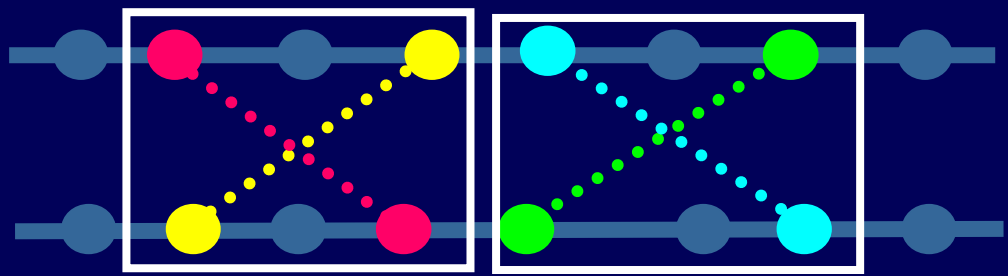
- Disjoint: A homologous gene pair should be a member of at most one cluster
- Isolated: The minimum distance *between* clusters should be larger than the maximum distance between homologous gene pairs *within* the cluster

Various types of constraints lead to overlapping (or nearby) clusters that cannot be merged

If we search for clusters with density ≥ ½:



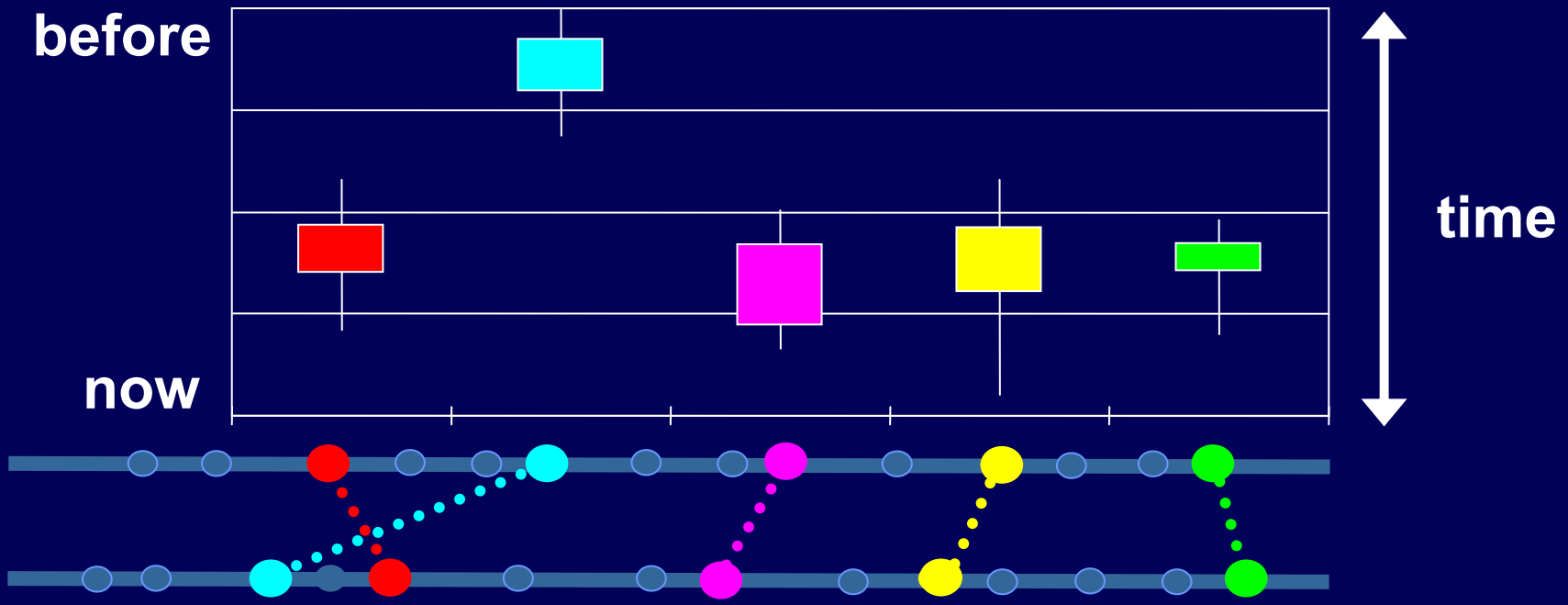
If we search for nested max-gap clusters, g=1:



Our Proposed Cluster Properties

- Symmetry
- Size
- Density
- Disjointness
- Isolation
- Order
- Orientation
- Nestedness
- Temporal Coherence

Temporal coherence



Divergence times of homologous pairs within a block should agree

Outline

- Introduction
- Brief overview of existing approaches
- Proposed properties for comparison
- My analysis of data: nested property
  - Many groups use a greedy, agglomerative search to find gene clusters
  - Does a greedy search have a large effect on the set of clusters identified in real data?

Data

	genes (1)	genes (2)	orthologs
E. coli & B. subtilis	4,108	4,245	1,315
Human & Mouse	22,216	25,383	14,768
Human & Chicken	22,216	17,709	10,338

} 2 pairwise genome comparisons

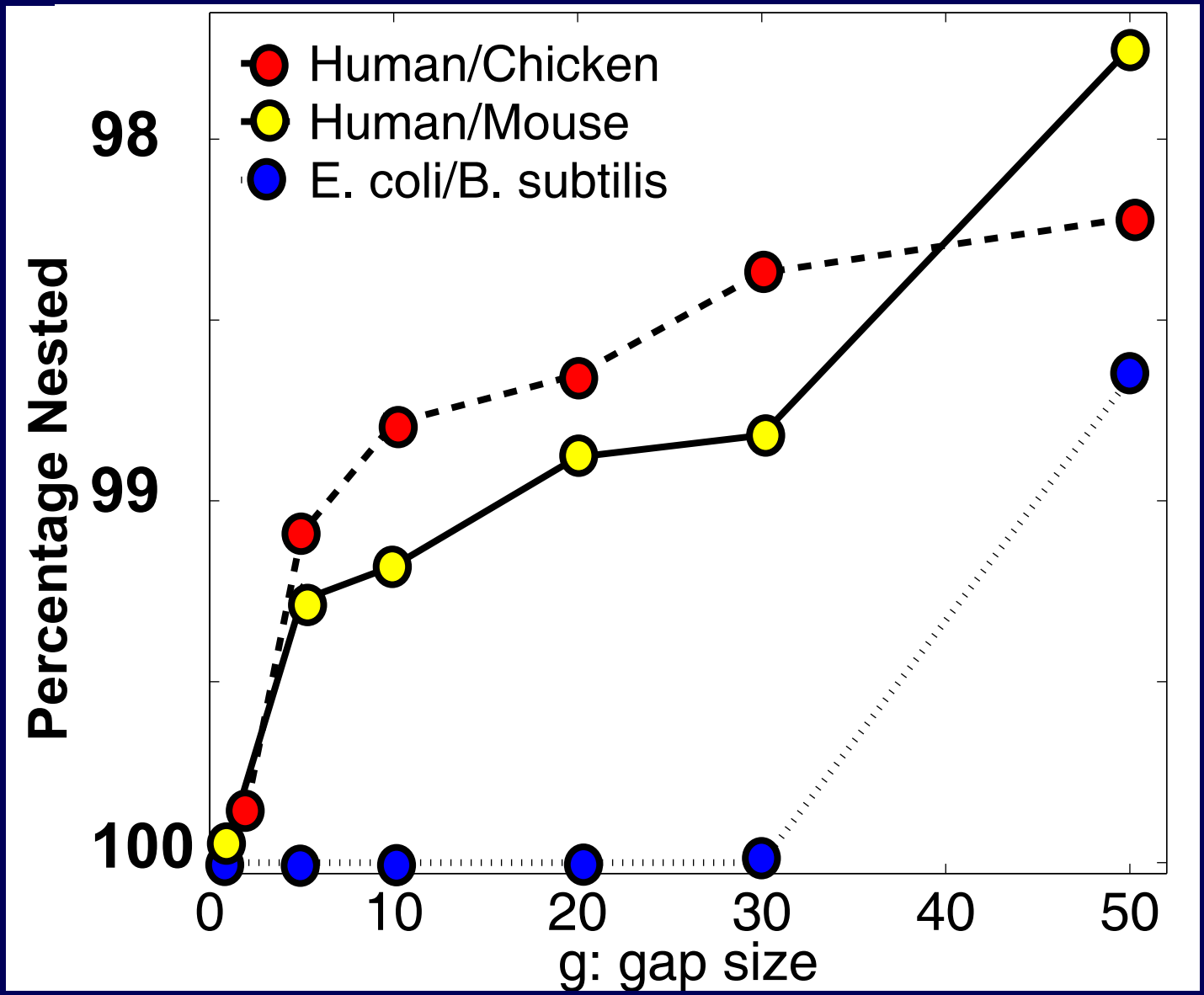
- Gene orthology data:
  - bacterial: GOLDIE database
    http://www.intellibiosoft.com/academic.html
  - eukaryotes: InParanoid database
    http://inparanoid.cgb.ki.se

Methods

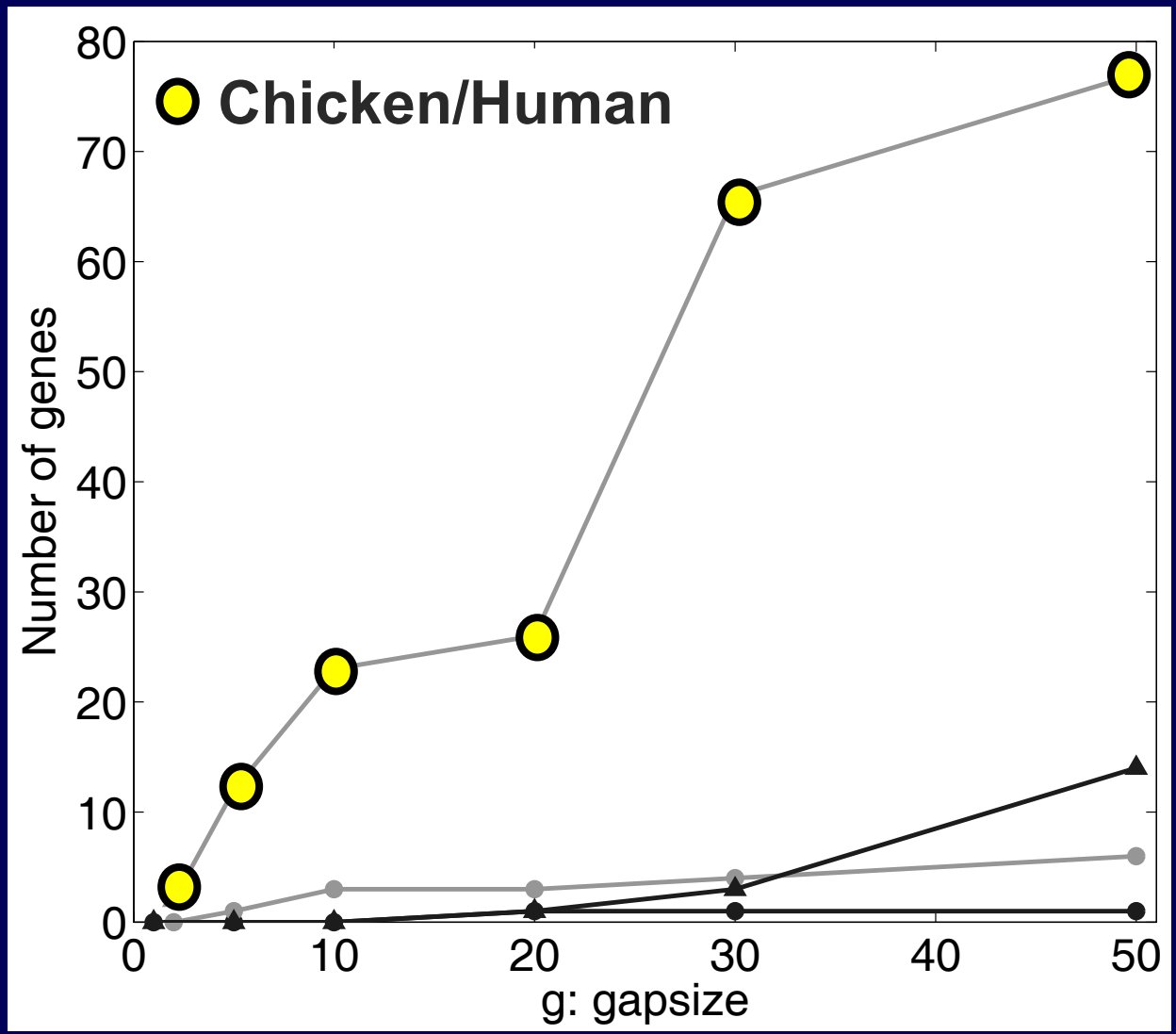
For each genome comparison and gap size:

- Maximal max-gap clusters
  - Gene Teams software
    http://.www-igm.univ-mlv.fr/~raffinot/geneteam.html
- Maximal *nested* max-gap clusters
  - simple greedy heuristic (no merging)

Percent of gene teams that are nested



Number of *genes* in some gene team of size 7 or greater that are not in any nested cluster of 7 or greater



Results

- For the datasets analyzed, a nestedness constraint does not appear too conservative

- However, we didn't survey a wide range of evolutionary distances
  - expect nestedness to decrease with evolutionary distance
  - open question: are there more rearranged datasets for which the proportion of nested clusters is much smaller?

Is nestedness desirable?

- A nestedness constraint:
  - offers a middle ground between no order constraints and strict order
- However, nestedness
  - provides no formal description of order constraints
  - is restrictive rather than descriptive
- We may instead prefer methods that
  - allow for parameterization of degree of disorder
  - consider order conservation in the statistical tests

Conclusion

- Proposed 9 properties to compare and evaluate methods for identifying gene clusters
- Illustrated cluster differences due to
  - cluster definition
  - search algorithm
  - statistics
- Incompatible Desiderata:
  - these properties are intuitively natural yet many are surprisingly difficult to satisfy with the same definition

Acknowledgements

- David Sankoff
- The Durand Lab
- Barbara Lazarus Women@IT Fellowship
- Sloan Foundation
- NHGRI, Packard Foundation

Discussion

- are our intuitions about clusters reasonable?
- which cluster properties are important or desirable?
- how can we quantitatively evaluate cluster definitions?
- what are the tradeoffs between methods?
- how can better definitions be designed?