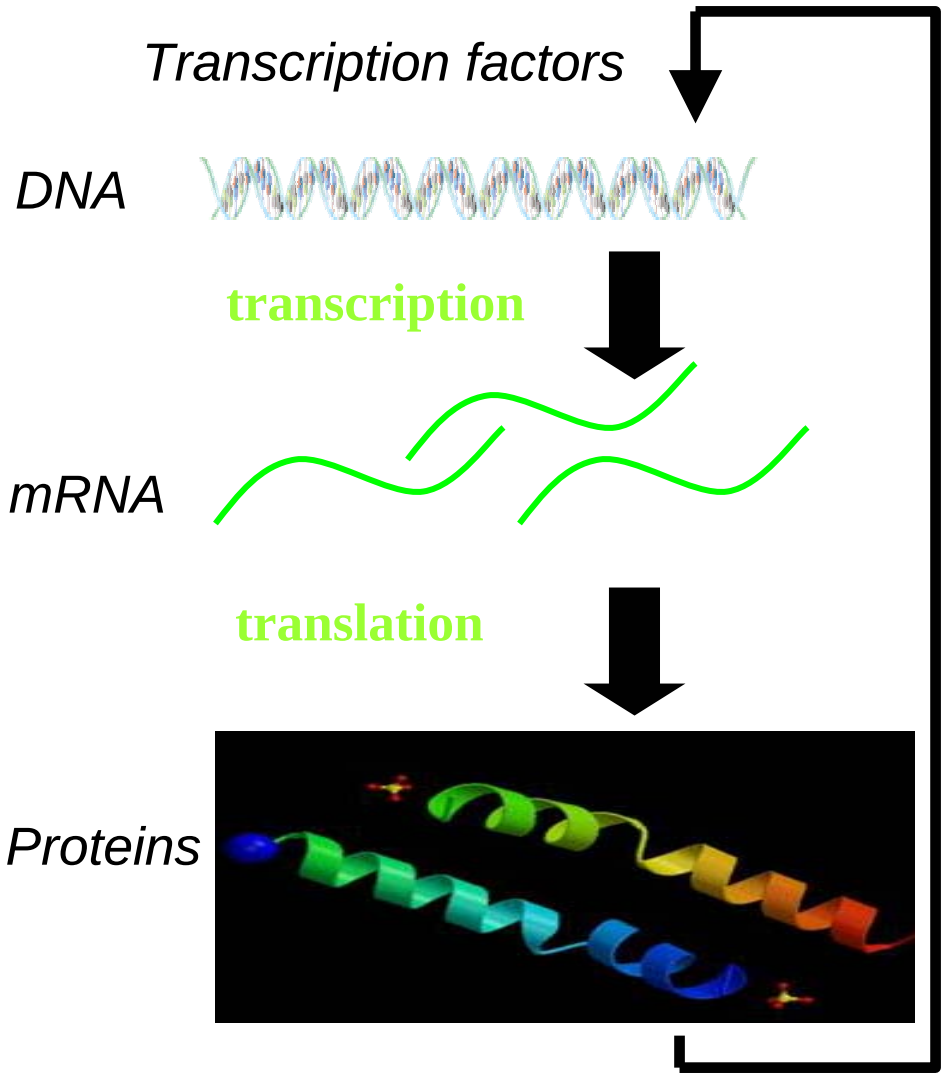
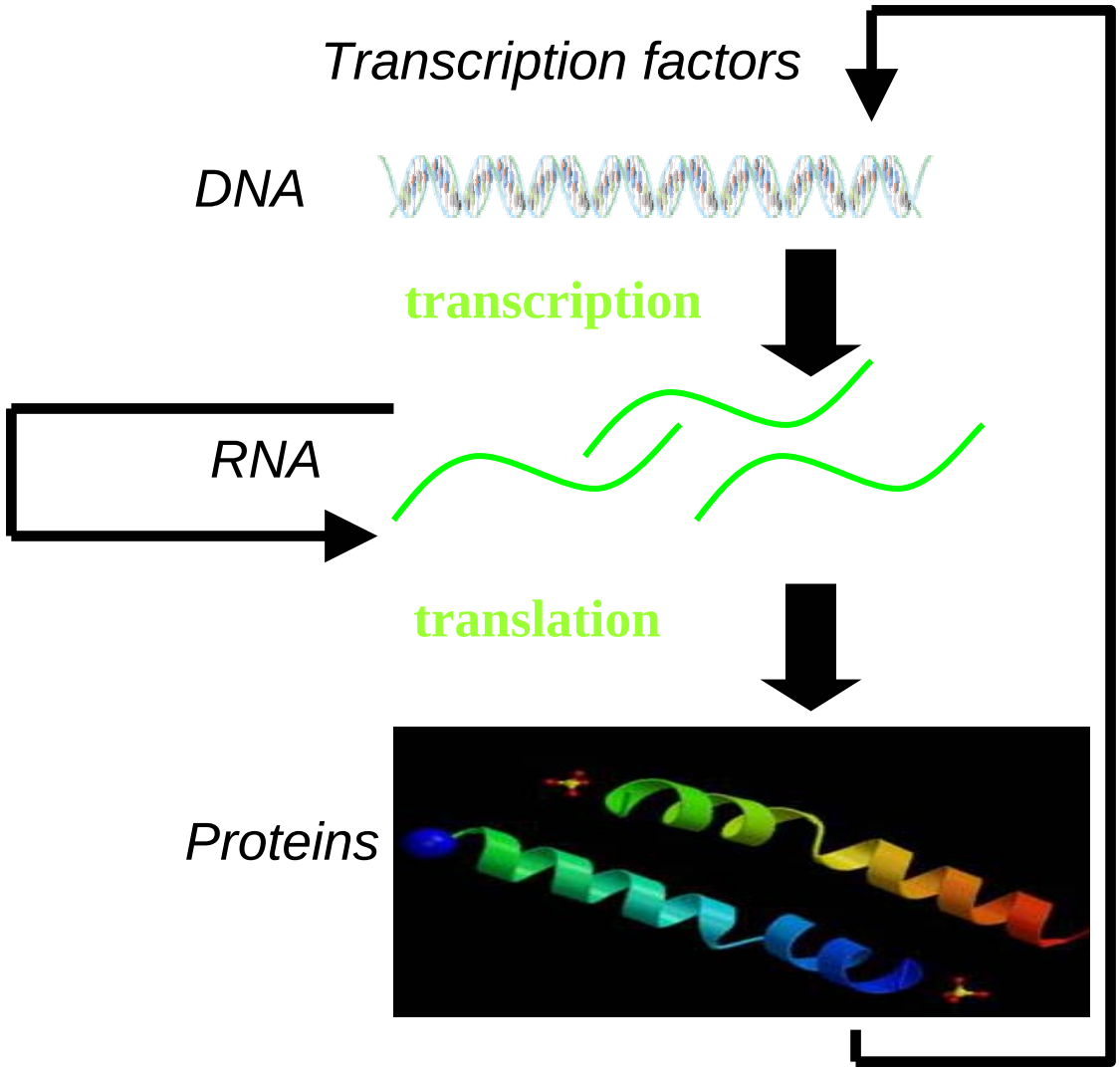
10-810: Advanced Algorithms and Models for Computational Biology

microRNA and Whole Genome Comparison

Central Dogma: 90s



Central Dogma: Updated

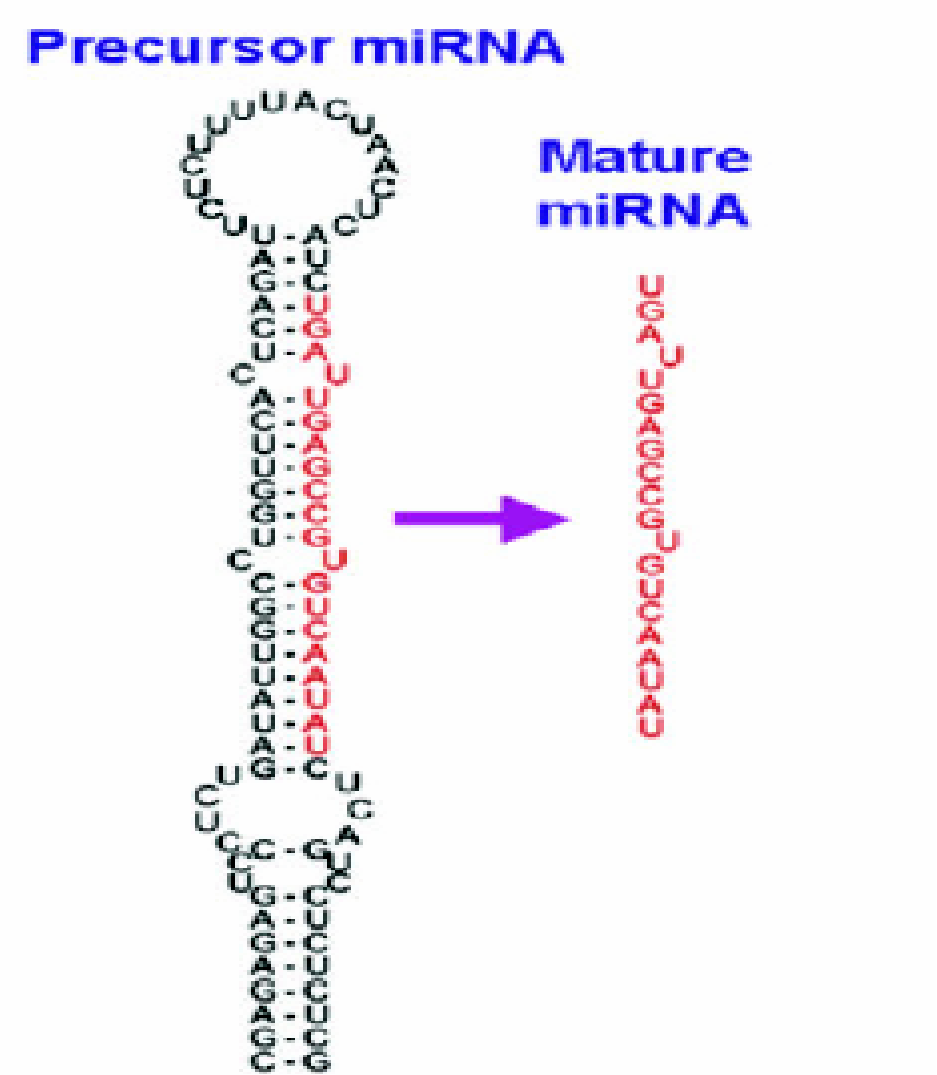


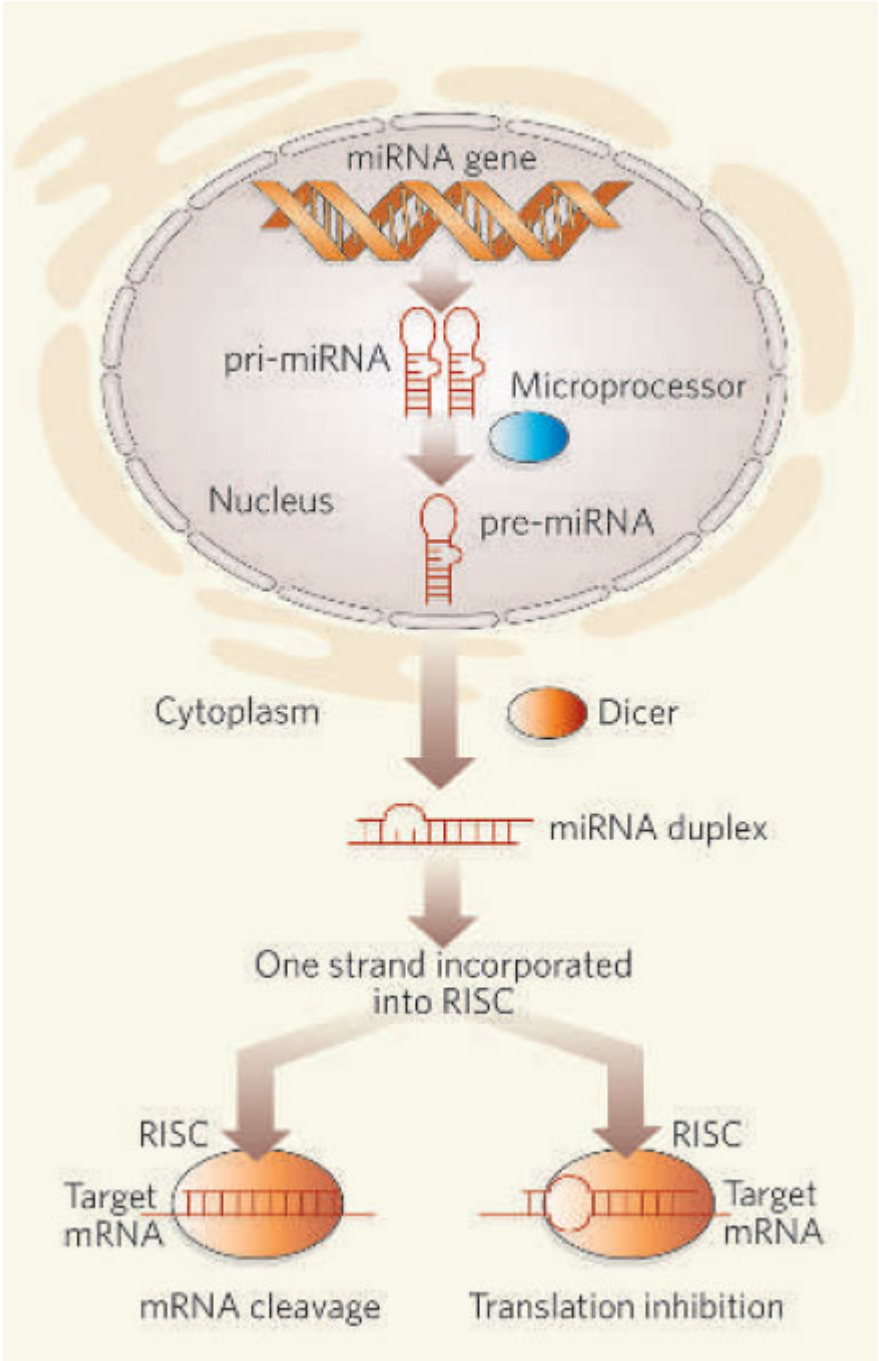
Regulatory non coding RNAs

- Two major types
  - Micro RNA (miRNA)
  - Silencing RNA (siRNA)
- Both are post transcriptional regulators
- Difference primarily in the way they regulate the mRNAs

miRNA

- Encoded as part of a longer RNA segment
- One arm used for binding to the regulated mRNA
- Follows a stem-loop structure
- Either binds to target mRNA resulting in cleavage or to 3' translated region (UTR) to prevent translation.




miRNA gene
pri-miRNA
Microprocessor
Nucleus
pre-miRNA
Cytoplasm
Dicer
miRNA duplex
One strand incorporated into RISC
RISC
Target mRNA
mRNA cleavage
RISC
Target mRNA
Translation inhibition

History

- First two miRNAs identified in early 90's in *c. elegans* (a small worm).
- More recently they were found to be conserved in multiple species.
- It is now believed that there are hundreds of miRNAs in higher organisms.
- Why is it useful to regulate on the mRNA level?

Identifying miRNA

- Given a complete genome we would like to identify the set of miRNAs (just as we do with genes).
- Problem: miRNAs are very short, and there are no clear rules (except for the stem-loop structure) for their sequence.
- This is very different from genes, for which much more structure information exists.
- How can we tackle this problem?

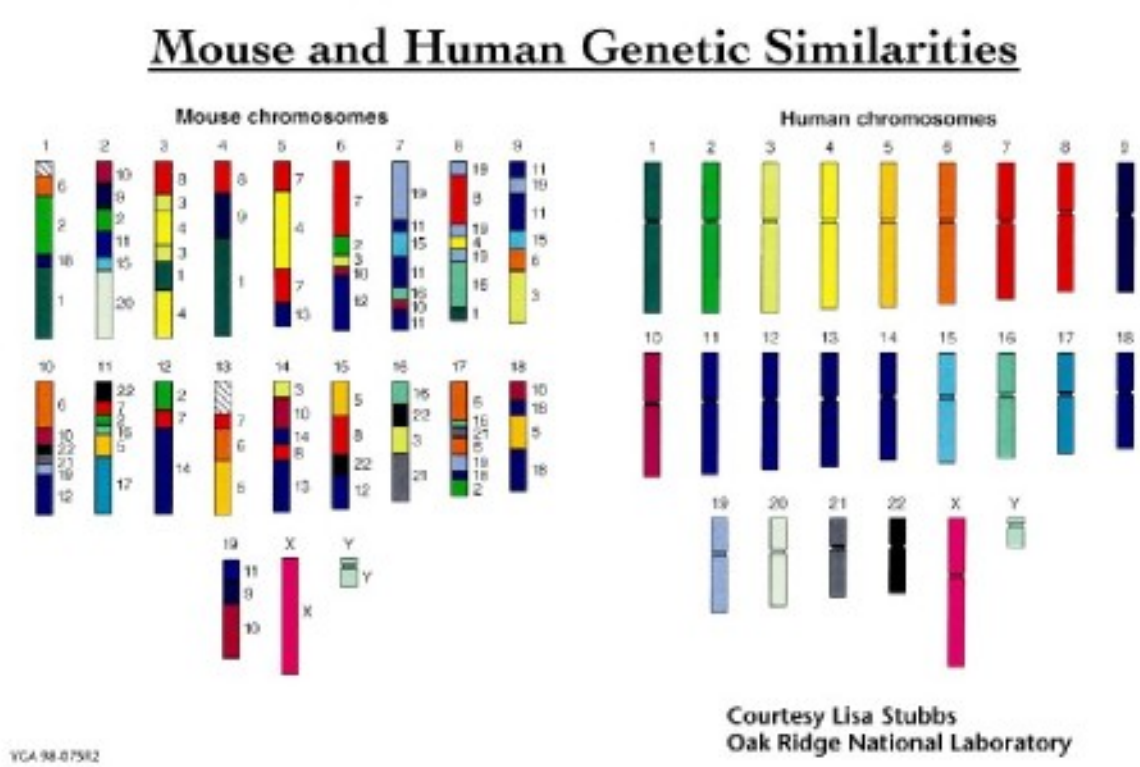
Whole genome comparison

Comparing genomes

- Recent advances in sequencing technologies are allowing researchers to sequence entire genomes very quickly.
- Lets assume that we know how to assemble a genome from the sequenced pieces.
- Given two genomes, X and Y, from closely related organisms, how do we determine a *global* alignment for them?
- Problems:

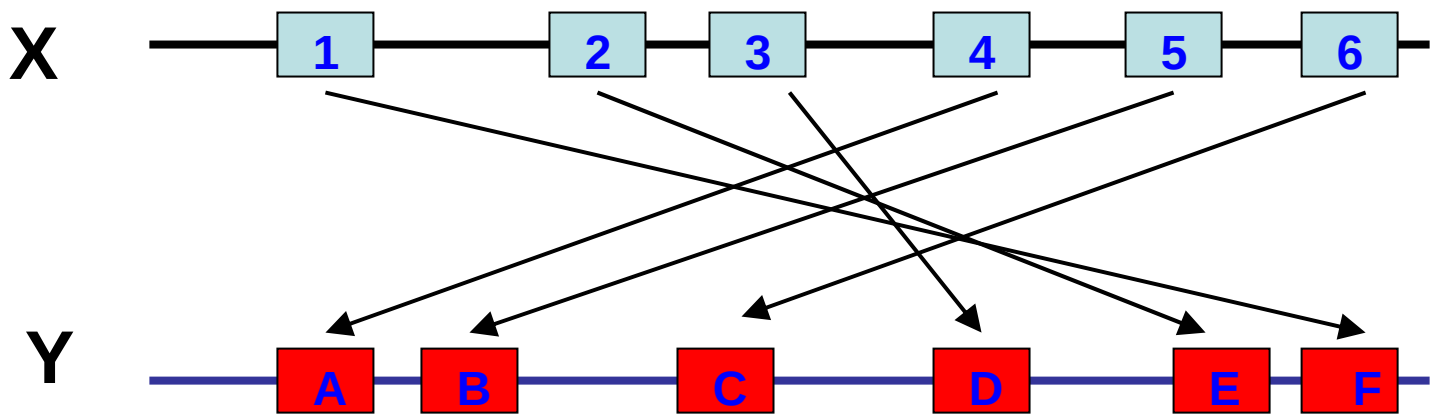
  - Mutations
  - Rearrangements
  - Duplications (even a whole genome duplication)
  - Etc.

Comparative genomics



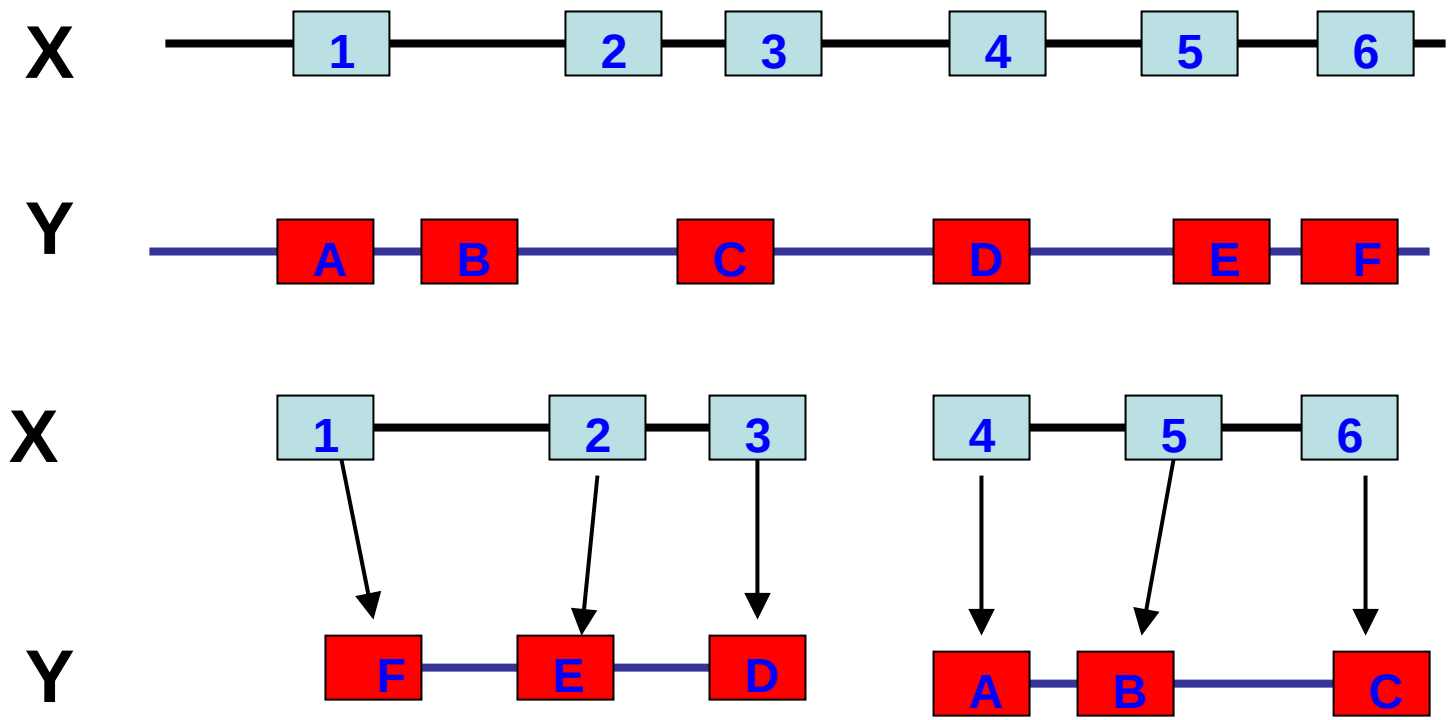
Anchors

- Key idea: Identify a set of anchors
- Determine relationships between anchors



Anchors

- Key idea: Identify a set of anchors
- Determine relationships between anchors
- Realign using the determined mapping

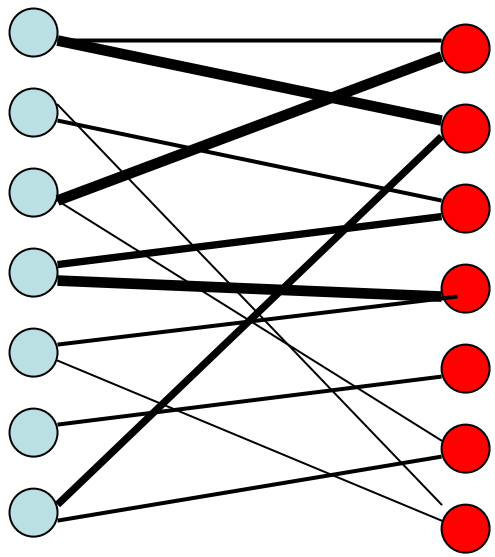


Genes as anchors

- Genes are natural candidates for anchors
- There is an evolutionary pressure to keep the gene sequence unchanged
- There are algorithms to identify the set of genes in an organism
- Key problem: determining the set of orthologs genes:
  - Duplications will lead to many to many relationships
  - Mutations are still possible
  - Paralogs will cause ambiguity

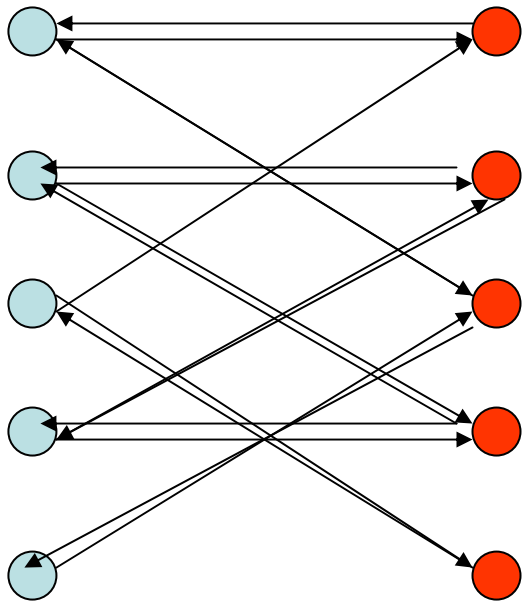
Solving the correspondence problems (kellis et al 2003)

- Use a (weighted) bi-partite graph
- Nodes correspond to genes
- Edges correspond to similarity
- Goal – resolve graph to obtain pairwise relationships and synteny blocks



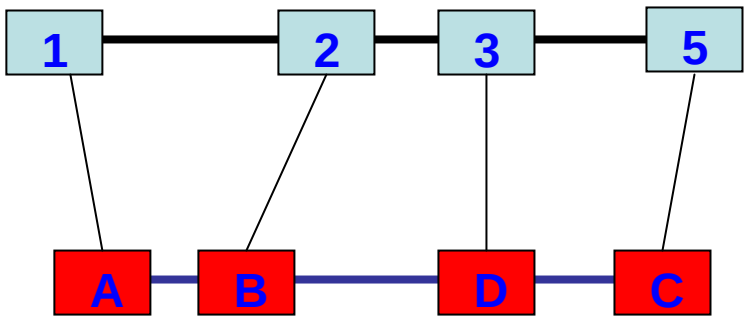
Step 1: Undirected to directed

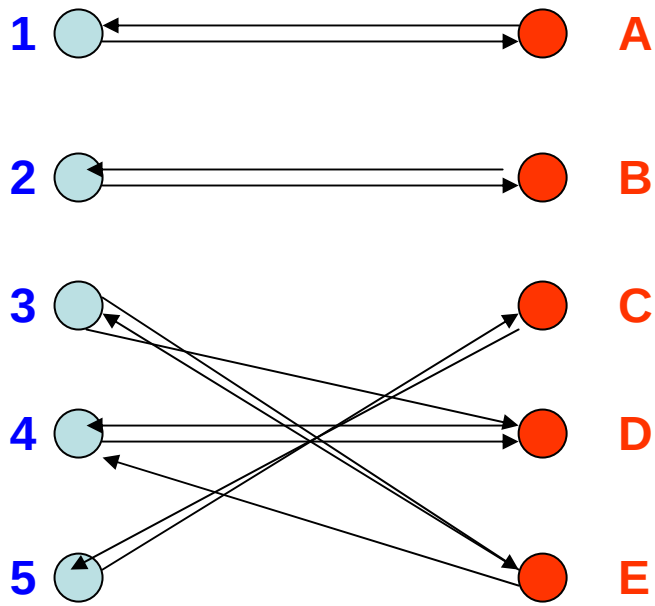
- Turn each edge to multiple edges



Step 2: Eliminate edges

- For each node, keep its *outgoing* edges only if they are at least 80% of the highest edge
- Use pairs to identify blocks

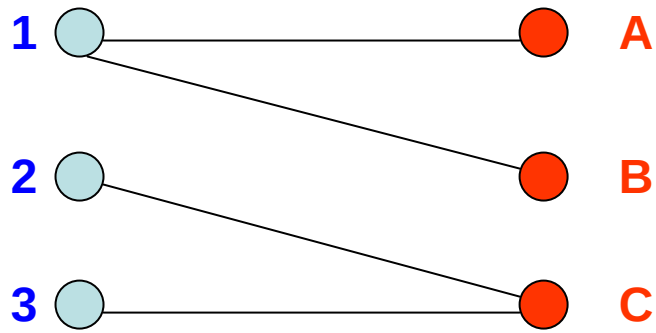


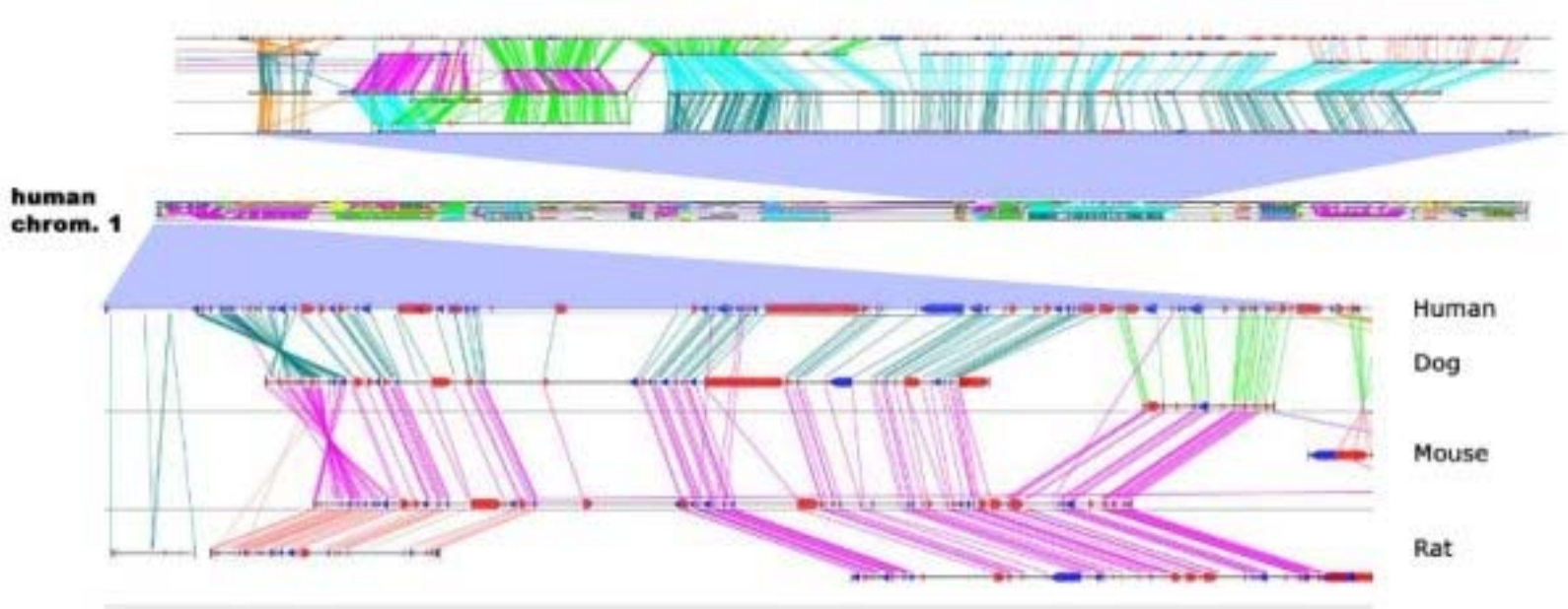


Step 3: Best Unambiguous Subset

- Edges that remain in the graph after step 2 are further pruned by removing all but the top outgoing edge(s) for each node
- The graph is then partitioned into connected components

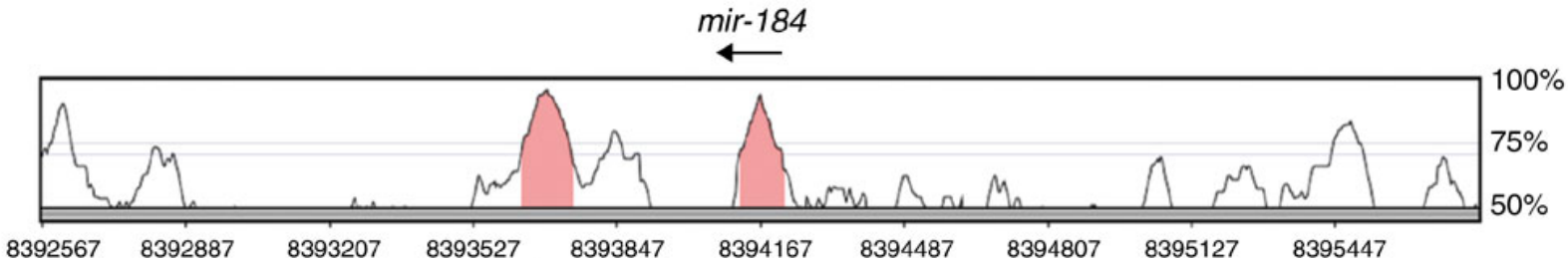
Again, many to one relationships are resolved based on synteny blocks




human
chrom. 1
Human
Dog
Mouse
Rat

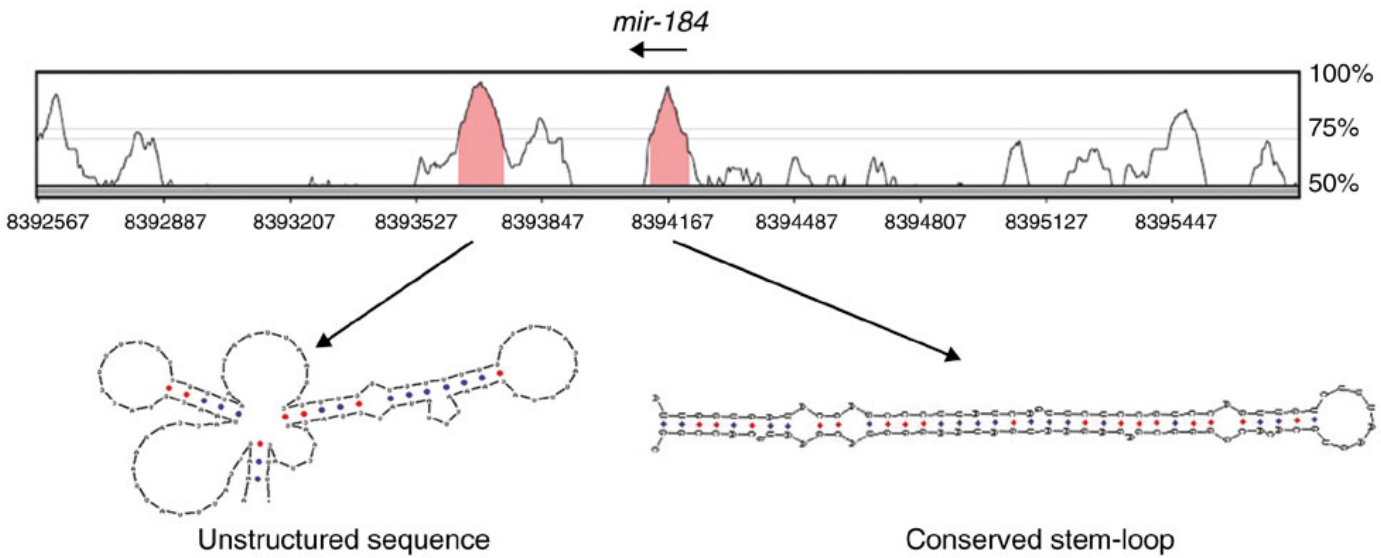
Back to microRNAs

- Given a whole genome alignment, we can now search for conserved segments, even if they are short.
- First step: identify conserved segments that fold to the correct structure (how can we tell?).



Scoring folds

- Conserved segments were folded using a RNA folding software
- Folds that exhibited non symmetric internal loops were panelized
- A final score was assigned to each fold



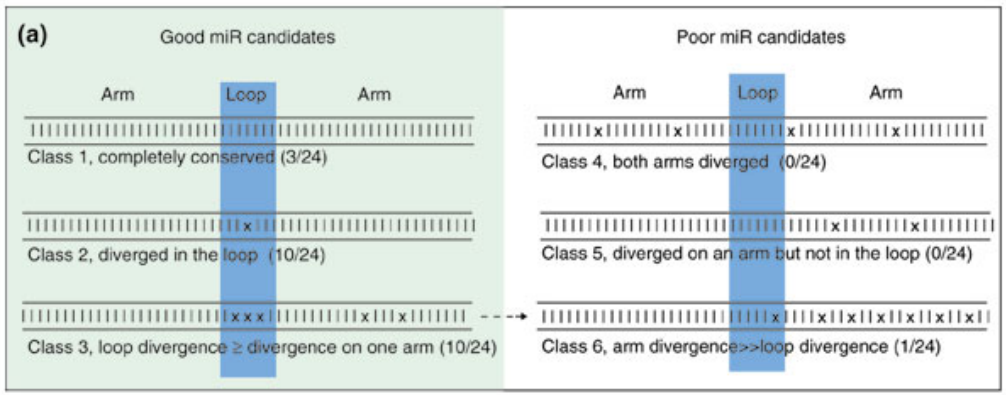
Conservation rate

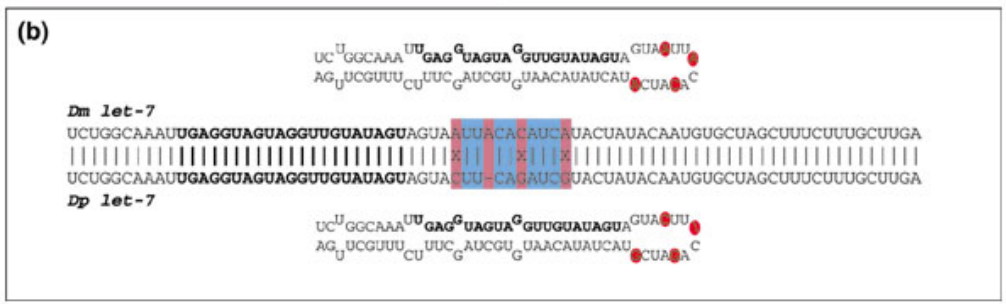
- The conservation of miRNAs in *Drosophila* (a fly) was studied using training data (a set of known miRNAs).
- Results indicated that the miRNAs where highly conserved, though the rate of conservation varied depending on the location

• Six classes of mutations were considered

• Three are present in real miRNAs and three are not

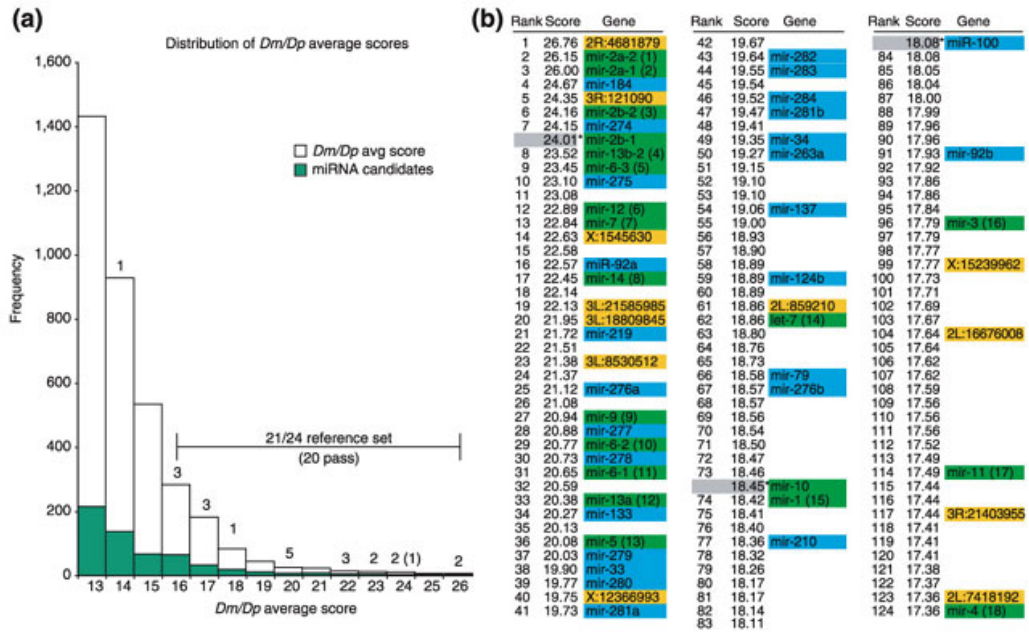
• Predicted sequences can be searched to eliminate those that do not agree with the three good classes





Putting it together

- After filtering for conservation classes, 200 high scoring candidate miRNA were left.
- These contained 18 of 24 training miRNAs and 182 predicted
- Most training data appeared in the top half



Experimental validation

- 20 of 27 (74%) predicted miRNA that were conserved in a third species were verified
- Only 4 out of 11 (36%) predicted miRNA that were **not** conserved in a third species were verified
- Authors claim that this is an upper value on the false positive rate since

  - Some miRNAs may only be expressed in certain conditions

  - Some may be expressed at very low levels

Predicting targets for miRNA

- Given a set of miRNA, the next question is to identify their targets.
- This is not a trivial task
- The binding may either be on the translated rna or on the 3’ UTR
- Direct comparison (with a folding software) leads to many false positives due to the short length of the miRNA
- A better strategy is to again rely on sequence conservation between organisms to identify these.
- Still, largely open problem

What you should know

- A revised look at the central dogma
- From pairwise sequence alignment ot whole genome alignment
- Much better to first look at the data and then devise the algorithm than the other way around