n

Towards a complete description of the microRNA complement of animal genomes

Gregor Obernosterer*

Bled 2004

*Part 3 of this presentation is missing due to my car-accident

Introduction

1. microRNA (miRNA) biogenesis
2. Estimation of miRNA complements
   - *Caenorhabditis elegans*
   - *Drosophila melanogaster*
3. Our proposed method for pre-miRNA estimation

Introduction

1. microRNA (miRNA) biogenesis
2. Estimation of miRNA complements
   - *Caenorhabditis elegans*
   - *Drosophila melanogaster*
3. Our proposed method for pre-miRNA estimation

Introduction

1. microRNA (miRNA) biogenesis
2. Estimation of miRNA complements
   - *Caenorhabditis elegans*
   - *Drosophila melanogaster*
3. Our proposed method for pre-miRNA estimation

Introduction

1. microRNA (miRNA) biogenesis
2. Estimation of miRNA complements
   - *Caenorhabditis elegans*
   - *Drosophila melanogaster*
3. Our proposed method for pre-miRNA estimation

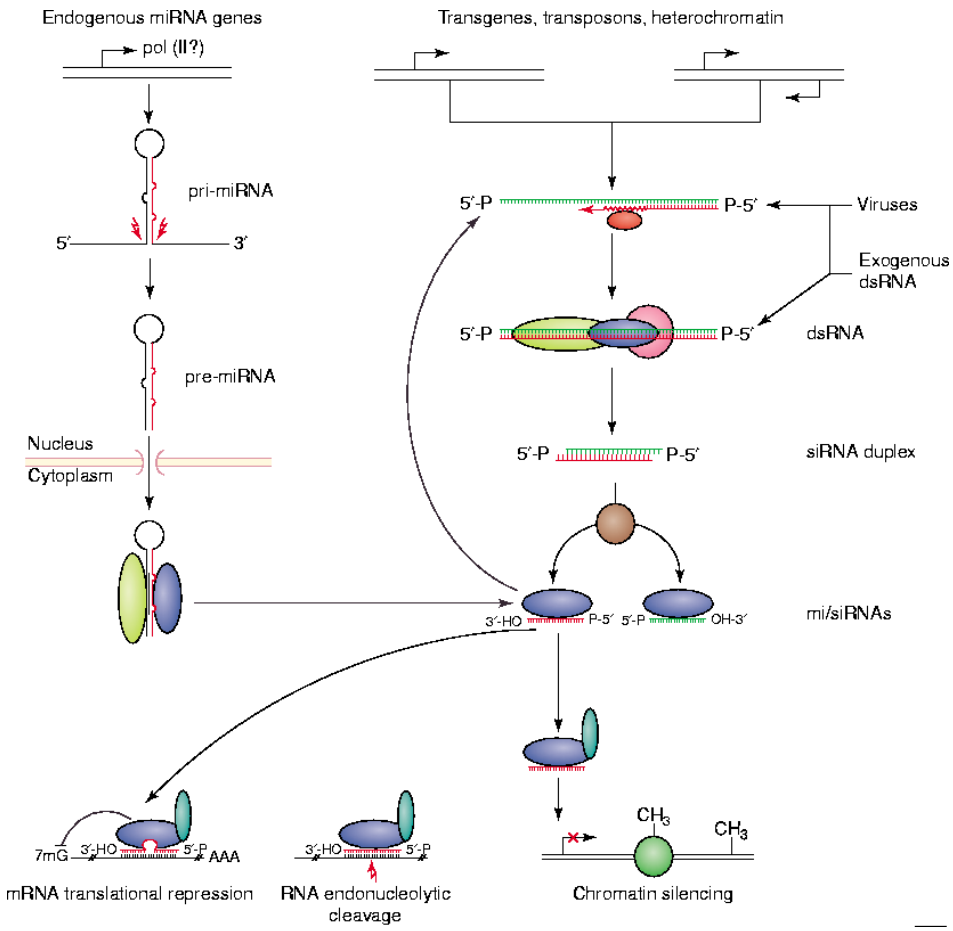
Introduction

1. microRNA (miRNA) biogenesis
2. Estimation of miRNA complements
   - *Caenorhabditis elegans*
   - *Drosophila melanogaster*
3. Our proposed method for pre-miRNA estimation

Introduction

1. microRNA (miRNA) biogenesis
2. Estimation of miRNA complements
   - *Caenorhabditis elegans*
   - *Drosophila melanogaster*
3. Our proposed method for pre-miRNA estimation

Proposed model of mi/siRNA-mediated gene expression regulation



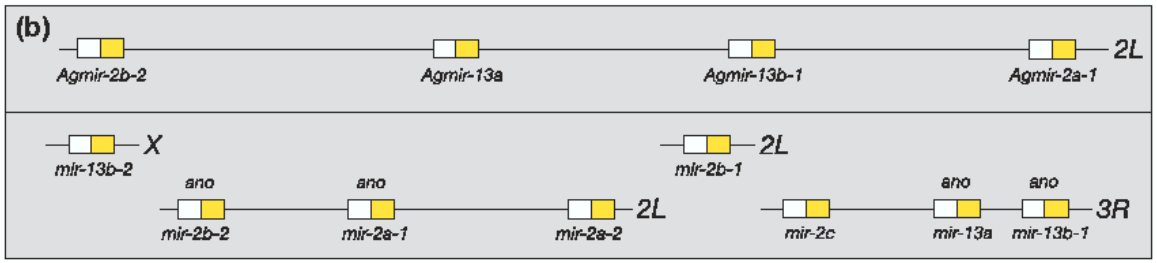
microRNAs versus siRNAs

1. Derived from an endogenous, structured transcript (pre-miRNA)
2. One miRNA accumulates
3. Evolutionary conserved
4. Usually located away from genes
5. Imperfect pairing blocks translation
6. Incorporated into miRNP
7. Regulate expression of genes encoded at another locus

1. Derived from extended dsRNA
2. Each dsRNA gives multiple siRNAs
3. Less conservation
4. Nearly complementary to target RNA (self-targeting)
5. Perfect pairing induces target RNA cleavage
6. Incorporated into RISC
7. Regulate the locus from which their sequence derives

Evolutionary Conservation of miRNAs

- miRNAs are evolutionary conserved even across phyla
- This suggests ancient and important roles for this class of regulators
- Observation: Found in multicellular plants and animals but not in unicellular eukaryotes
- Question: How many of these tiny regulators are hidden in animal genomes?



Cloning versus computational approaches

- Cloning endogenous RNA (18-25nt) has proven to be powerful
- More than 200 miRNA-coding genes have been identified
- Limitations
  1. Limited transcript abundance
  2. miRNAs at low expression levels might not be found
  3. Background from other small RNAs
- Alternative approach: Computational strategies
  1. Hairpin-like structures residing in intergenic or intronic sequences are identified
  2. The identified hairpin-set is refined by applying a series of structural filters
  3. Sequence conservation filters are applied for further refinements
- Successful in identifying most cloned miRNAs and identifictation of new miRNAs
  - ⋆ Northern-Blots, PCR-based assays

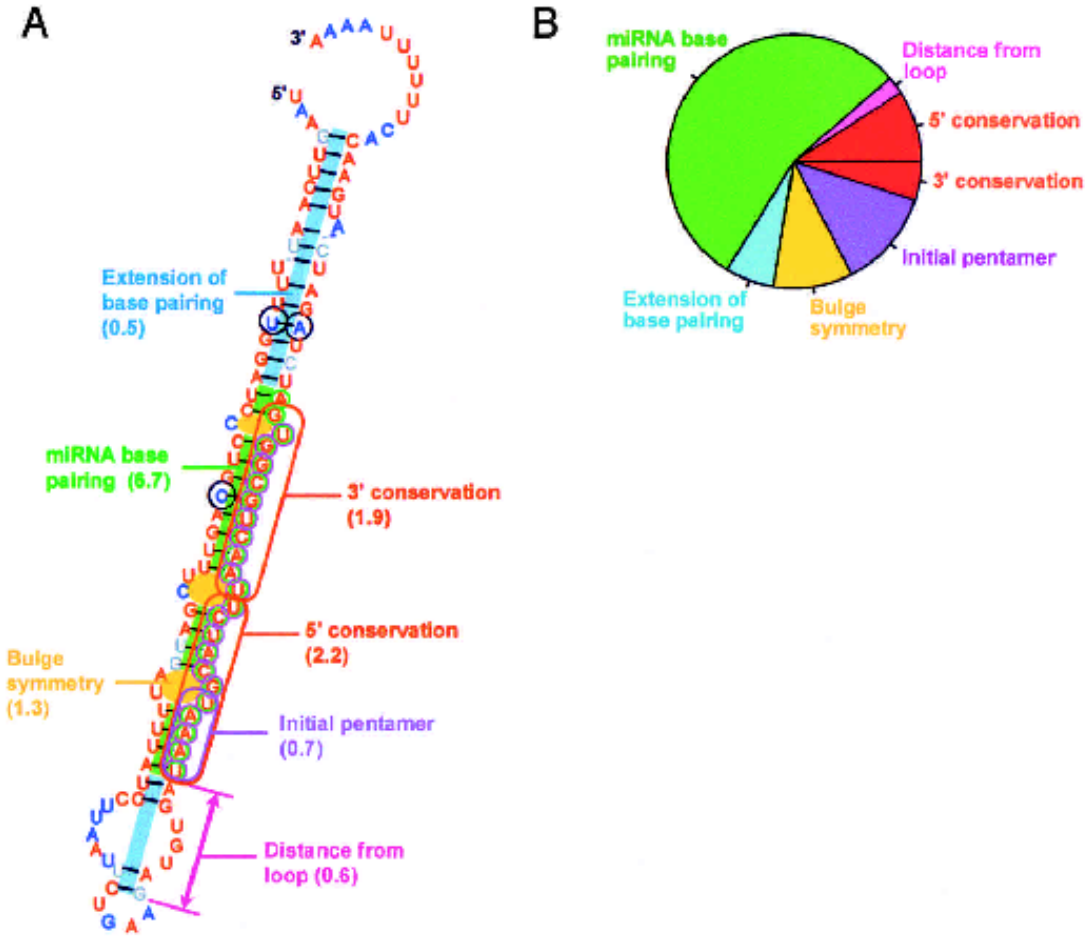
The microRNAs of *Caenorhabditis elegans* - MiRscan [1]

- Comparative search for conserved miRNA like hairpins

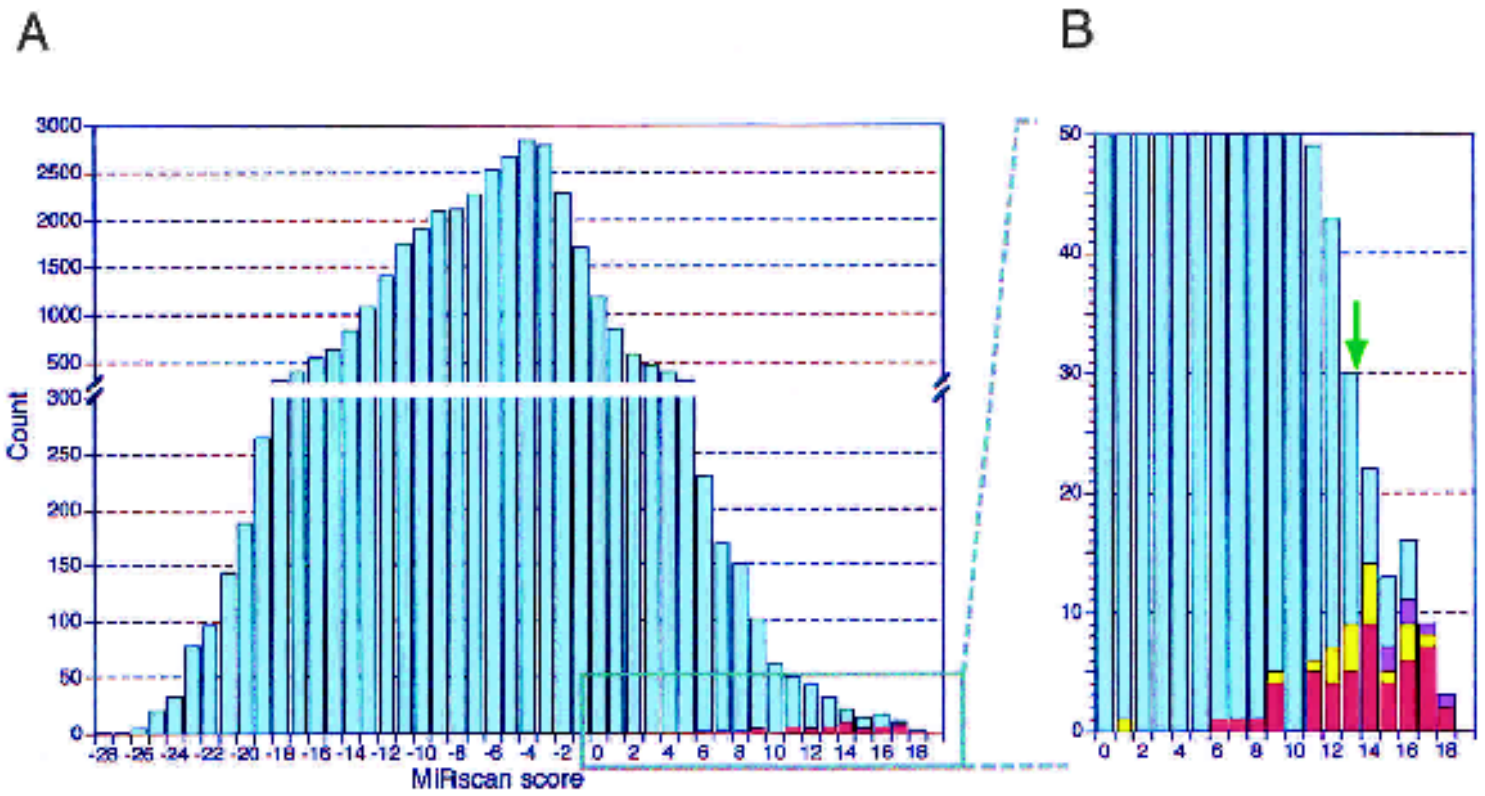
1. Scan for hairpin structures that were conserved in *C. briggsae* and *C. elegans* - 36,000 hairpins
2. 50 conserved miRNA genes served as training set for MiRscan
   - base pairing of the miRNA portion of the fold-back
   - base pairing of the rest of the fold-back
   - stringent sequence conservation in the 5' half
   - sequence bias in the first five bases of the miRNA (Uracil)
   - tendendcy toward having symmetric bulges
   - presence of two to nine consensus base pairs between the miRNA and the terminal loop region

[1] L.P. Lim, D.P. Bartel et al., Gens and Development **17**:991-1008, 2003

MiRscan scoring criteria



Distribution of MiRscan scores



MiRscan accurarcy

- Specificity: $\geq$0.70 at a sensitivity that detects half of the known *C. elegans* miRNAs
- Accuracy: Sufficient to identify new genes and obtain an upper bound on the total number of miRNAs
- **However**, not reliable to identify all the conserved miRNA genes
- Accuracy compared to other general methods to identify ncRNAs
  - ⋆ as high as methods to identify ncRNAs[2] in bacteria
  - ⋆ lower than that of algorithms that detect protein-coding genes, tRNAs or snoRNAs[3]

[2] Argaman et al. 2001; Rivas et al. 2001; Wasserman et al. 2001

[3] Lowe and Eddy 1997, 1999; Burge and Karlin 1998

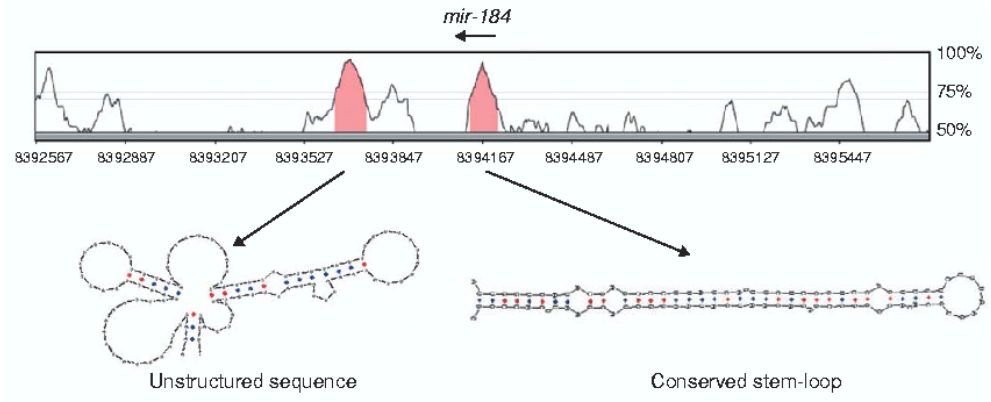
The microRNAs of *Drosphila melanogaster* - miRseeker[4]

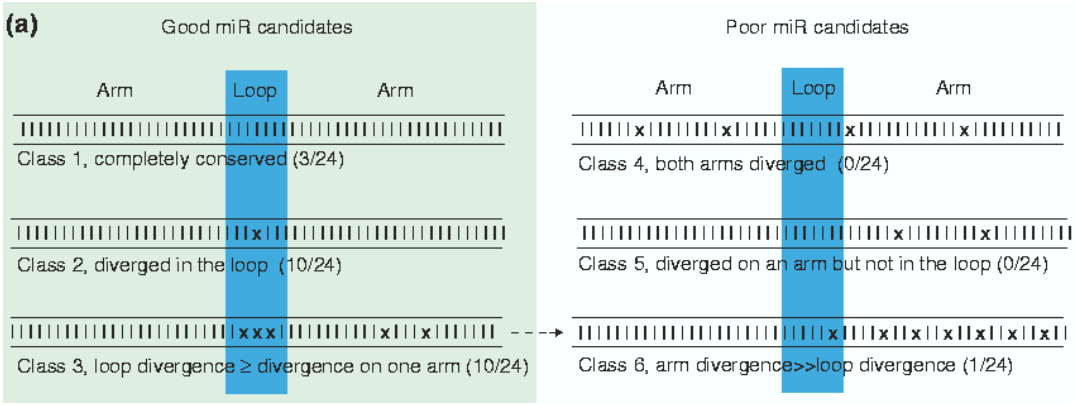
- Comparative search for conserved miRNA like hairpins

1. Scan for hairpin structures that were conserved in *D. melanogaster* and *D. pseudoobscura* - 436,000 100bp regions in 118,000 super-regions
2. 24 conserved miRNA genes (let-7, 21 by Lagos-Quintana, mir-125, mir-2c) served as training set for miRseeker
3. Assess the pattern of nucleotide convergence by aligning the 24 paors of orthologous *Drosophila* pre-miRNA sequences
   - Class 1: Completely conserved
   - Class 2: Diverged in the loop
   - Class 3: Loop divergence $\geq$ divergence on one arm
   - Class 4: Both arms diverged
   - Class 5: Diverged on an arm but not in the loop
   - Class 6: Arm divergence $\geq$ loop divergence

[4] E.C. Lai, P. Tomancak et al., Genome Biology **4**:R42, 2003

Classification of conserved stem-loop sequences and VISTA plots of globally aligned sequence from *D. melanogaster* and *D. pseudoobscura*





Computational prediction of *Drosophila* miRNA genes using miRseeker

- Extraction of candidate, conserved, 'nongenic' *Drosophila* sequences
- Identification of conserved stem-loops and evaluation of their quality
  - ⋆ length of the longest helical arm
  - ⋆ the free energy of this arm with ΔG $\leq$-23.0
  - ⋆ the presence of asymmetric loops and bulged nucleotides was further penalized
- Evaluation of the divergence pattern in conserved stem-loops
  - ⋆ Class 1 to 3 are referred as good-candidates
  - ⋆ Class 4 to 6 are poor candidates

Drosophilid genomes contain around 110 microRNA genes

- Catalogued 32 newly verified miRNAs
- Estimation of about 110 possible miRNA genes
- Unique aspect: Assessment of the pattern of nucleotide divergence within miRNA precursors

Conclusio

- Problem with sorting out new miRNA genes from random sequences that can form plausible hairpins
- Only 50-75% of previously validated miRNAs were among the 'high confidentiality' set
- Identification seems to be hampered by our limited knowledge of sequence and structural features that distinguish them from background 'hits' in the genome
- Sequencing additional vertebrate, worm and insect genomes is likely to be a powerful resource for improving computational prediction methods
- Computational methods only allow the identification of genes that resemble those in the training set