# Decoding the long-term history of species interrelationships using mitochondrial genomes

Jack Schwartz

New York University

(Joint Work with Michael Wigler,

Cold Spring Harbor Laboratory)

Mitochondria

- "Mitochondria are the powerhouses of the cell. They take in fuel molecules derived from sugars and fats, harvest the energy in their bonds with the aid of oxygen, and spit out ATP, the universal energy carrier needed throughout the cell to fuel the energy-hungry reactions of life."
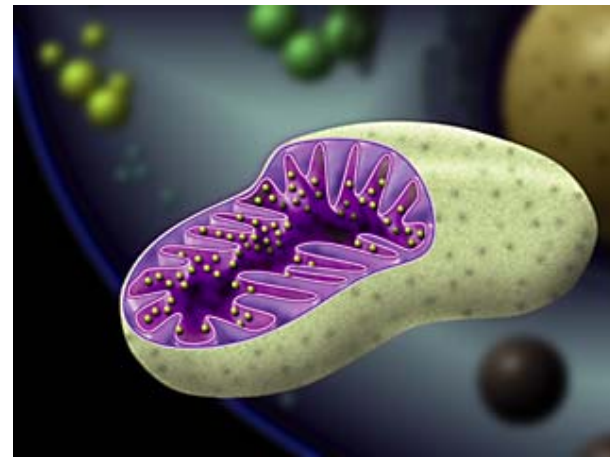

- Mitochondria are germ-sized
  - 100-600 mitochondria per cell
  - Have their own DNA 'cache' but depend largely on imports from host cell
  - Inheritance is purely maternal

Mitochondrial DNA holds a standard set of 37 genes over a very wide range of species

- **Excepting protozooans and plants (ignored in this talk)**

- **22 'transfer' RNAs (tRNAs) including two duplicated**
  - uses slightly non-standard genetic code
- **2 'ribosomal' RNAs (rRNAs)**
- **13 genes for proteins (the key proteins involved in processing of oxygen)**
  - Cytochrome oxidases(1-4); NADH synthases (1-7); ATP synthases (2)

- **Focusing on this 'small machine' (vs. the 50,000 gene 'large machine' which is full DNA) eases understanding of inter-species relationships**

  - The genome is 16KB rather than many MB or GB
  - The genes and their positions are known
  - Complete genomes are available for hundreds rather than dozens of species

- **These advantages make use of mitochondrial DNA common in studies of phylogeny.**

Histogram of mitochondrial genome lengths
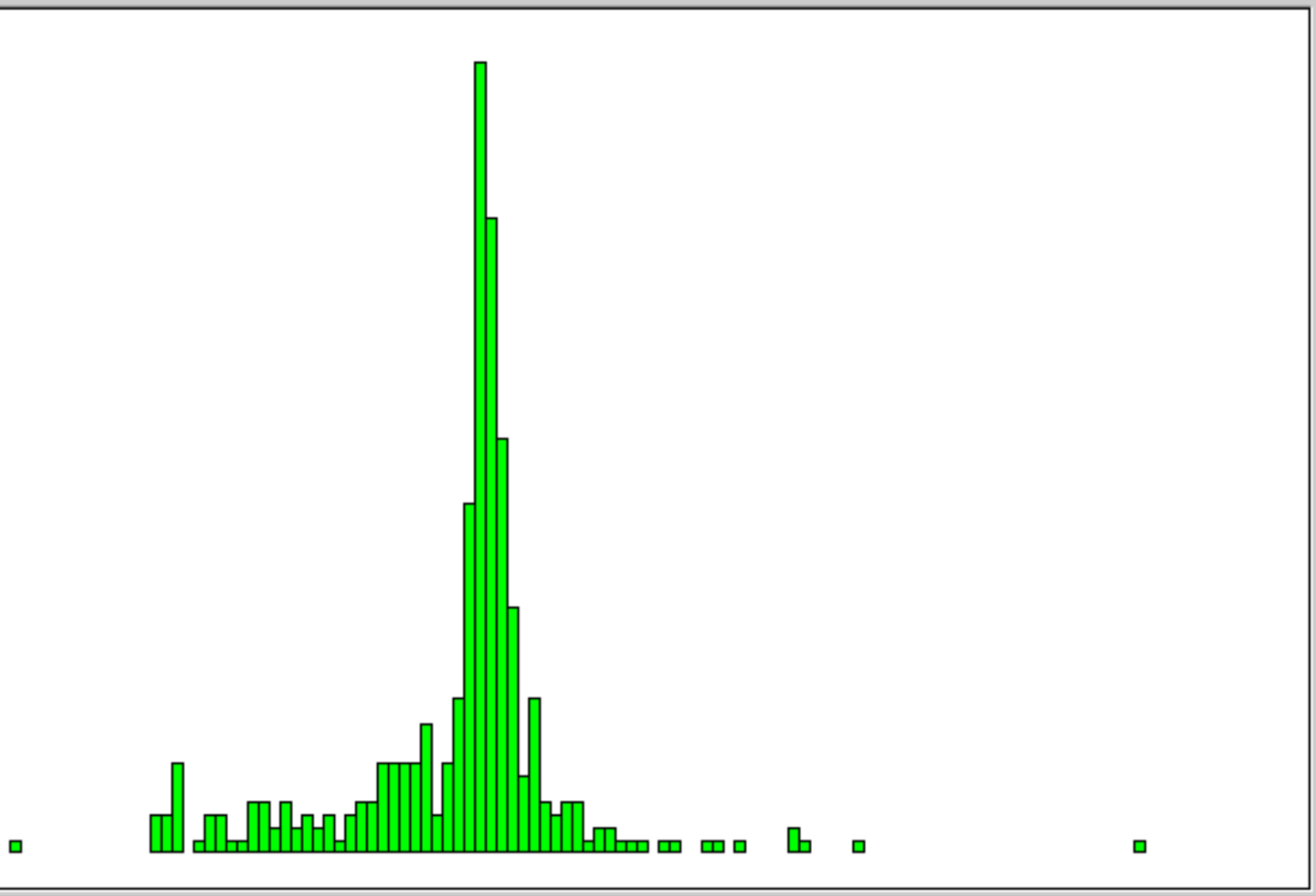
But a careful approach is needed if clear relationships are to be seen

- **It is difficult to derive unambiguous conclusions from detailed base data for species differing widely.**

- **A clearer (though much coarser) picture results if only the order of genes on the mitochondrial DNA is considered, rather than the internal detail of the individual genes.**
  - **Inspection at 3 levels is possible:**
    - **use 15 'large' genes only (ignore 22 tRNAs, which are about 70 bases long, use others, which are several hundred long).**
    - **Use all 37 mitochondrial genes**
    - **Also use flags indicating strand containing gene.**

- **Method: assign (arbitrary) 1-character codes to genes, represent order by a string**

  Anopheles   pEBFSJPrHqQTRszSmL6V2IDMo&C#eLfKAbagY
  Human   P2V6LmIDMo&BFC#eSAfKbagYpEQqHSLrsJzTR

- **Begin inspection with these 37 character codes**
  - **But look first at the surrounding history**

Archeology timeline - Dates in million years ago

- **3800 (start of) Archean**   **Earliest microbial fossils**
- **2500 Paleoproterozic**   **Eucaryotes, microbial mats**
- **1600 Mesoproterozoic**   **Algae**
- **900 Neoproterozoic**   **Multicelled Animals**
-
-    **650 Vendian**   **Sponges, Worms**
-    **543 Cambrian**   **Trilobites, jawless fish, vertebrates**
-    **510 Ordivician**   **Fish**
-    **443 Silurian**   **Life on land, land plants**
-    **417 Devonian**   **Land plants, land insects, salamanders**
-    **354 Carboniferous**   **Egg-laying tetrapods, centipedes, scorpions**
-    **290 Permian**   **Pre-mammals, sharks (Extinction ***)**
-    **245 Triassic**   **Early dinosaurs**
-    **205 Jurassic**   **Dinosaurs and other reptiles**
-    **146 Cretaceous**   **Late dinosaurs (Extinction ***)**
-    **65 Cenozoic**   **Modern period**

Pikaia
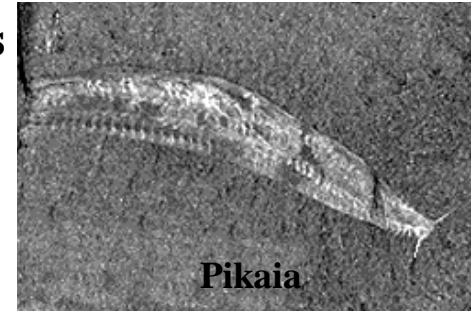
650 Million Years ago- Vendian
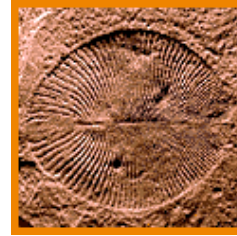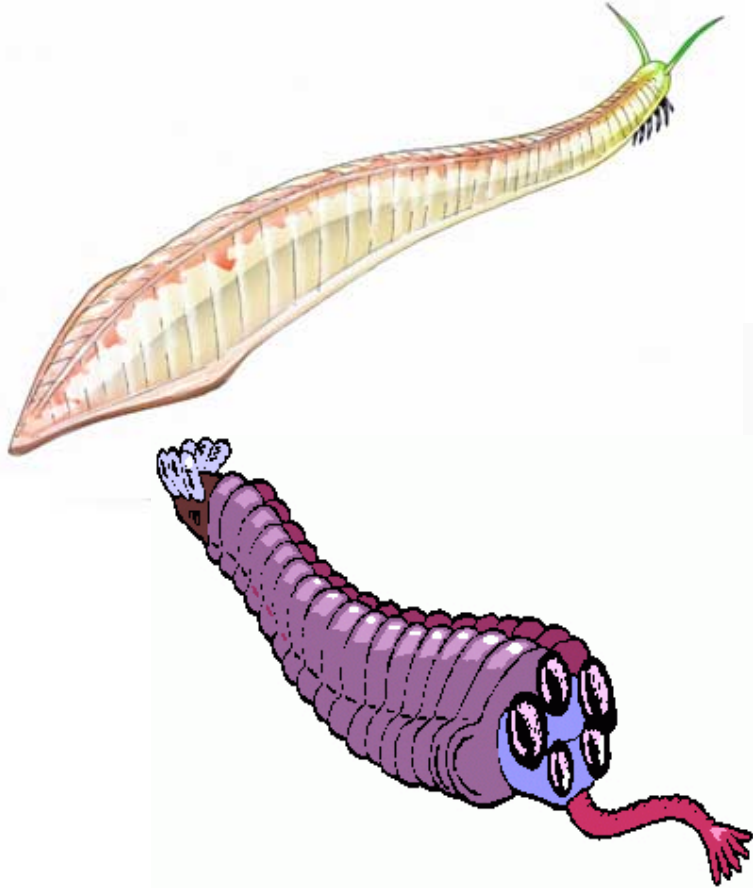
510 Million Years ago- Ordovician

heterocercal tail

mouth

eye

nostril

gill opening

small scales cover ventral surface of head

ventrolateral ridge

flipper
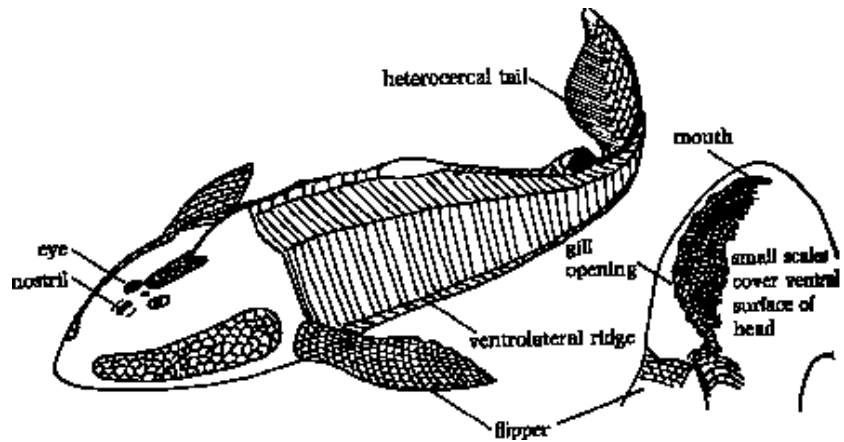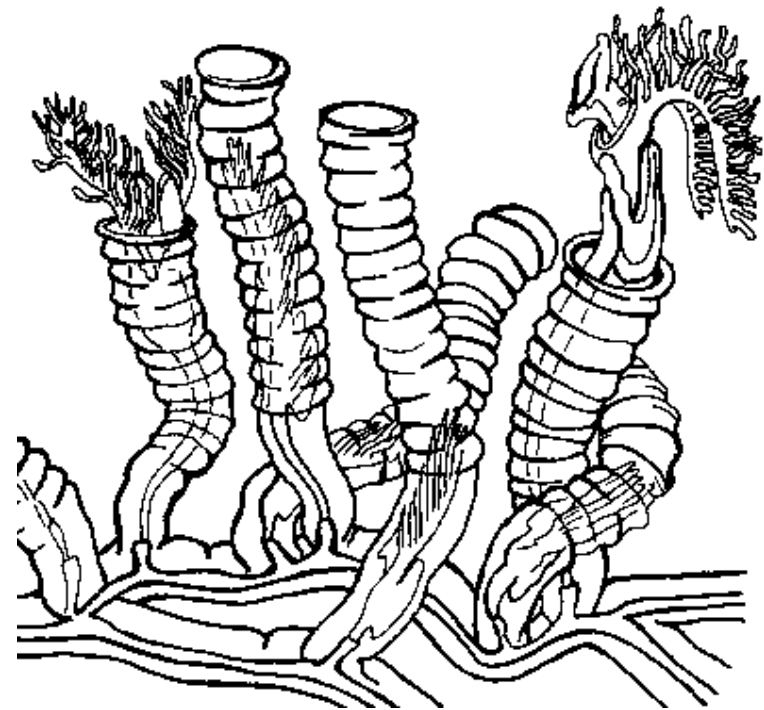
205 Million Years ago - Jurassic
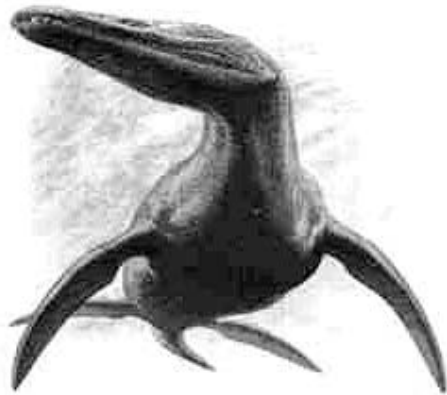
65 Million Years ago - Cenozoic

- **Meanwhile the continents have been floating around…**

First results: eyeball, establish standard templates, compare to standard templates, highlighting differences

- **Vertebrate**   P2V6LmIDMo&BFC#eSAfKbagYpEQqHSLrsJzTR

- **Arthropod**   pBEFSJPrHqQTRszSmL6V2IDMo&C#eLfKAbagY

- **Gastropod**   QzACPf#&YHDLbFaEJ2MpSSqTgIoKeV6LRBsrm

- **Echinoderm**   P2JTRDFLB&CVMA#YLmIo6eEQfKbagSpqHSrsz

- **Platyhelminth**   gHzQqDPMaoVBAmFRIKpS&eT6C2fJs#LSLErY

- **Nematode**   RVsQ&J2SF#maKLSoIEDPzLgTqeCMAYfH6prB

- **Roundworm**   Q2#mPaIYfH6pCSRAVrJSoTqe&sEDzLgKBLFM

- **Schistosome**   MaoBAmFRIKp&VSeT6C2fsL#SLErYgJHzQqDP

- **Seasquirts**   HEDgK6IpJVqPCY#LSerAML2MT&SsYFaoBfzmRQ

- *Note: protozooans, fungi, and plants, which are very different, are ignored*

First comparisons

- **133 odobenus_walrus** ....
- **134 Monocentris_pineconefish** ....
- **135 human** ....
- **137 Manis_pangolin** ....
- **139 Neoceratodus_lungfish** ....
- **143 Gymnothorax_eel** ....

- **46 triops_shrimp** ------------------------------------
- **68 panulirus_lobster** ------------------------------------
- **77 crioceris_beetle** ------------------------------------
- **87 Cochliomyia_screw_worm** ------------------------------------
- **83 Penaeus_prawn** ------------------------------------

- **These identities, and hence the templates they represent, clearly antedate the emergence on land, roughly 400 million years ago.**

Finer comparisons

- 50 Eudromia_crested_tinamou ....zTRsJ
- 146 Struthio_ostrich_bird ....zTRsJ
- 151 Aythya_redhead_bird ....zTRsJ
- 183 Coturnix_quail_bird ....zTRsJ
- 185 Coturnix_japan_quail ....zTRsJ
- 187 Tinamus_extinct_bird ....zTRsJ
- 188 Rhea_bird ....zTRsJ
- 192 Dromaius_emu_bird ....zTRsJ
- 194 Anomalopteryx_moa ....zTRsJ
- 198 Anser_goose_bird ....…zTRsJ

- 223 isoodon_bandicoot ....BC&F........................
- 231 Macropus_wallaroo ....BC&F........................
- 236 Wombat ....BC&F........................
- 247 Didelphis_opossum ....BC&F........................
- 252 Trichosurus_possum ....BC&F........................

- *The originals are sJzTR and &BFC*

Finer comparisons

- **166 alligator** ....SH.......
- **200 alligator_chinese** ....Yg....SH.......
- **267 caiman** ....SH.......

- **Lepidoptera among arthropods**

- **29 Ostrinia_european_borer** -------------------MID------------
- **30 Ostrinia_asian_borer** -------------------MID------------
- **59 antherea_moth** -------------------MID------------
- **61 bombyx_worm** -------------------MID------------
- **64 bombyx_xiafang_moth** -------------------MID------------
- **80 bombyx_silkworm** -------------------MID------------

- **24 orthinodorus_tick** -----------------L6V2IDMo&C#e-------
- **31 ixodes_hedgehog_tick** ----------------L6V2IDMo&C#e-------
- **32 Ixodes_tiger_tick** ----------------L6V2IDMo&C#e-------
- **42 Limulus_crab** ----------------L6V2IDMo&C#e-------
- **65 Lithobius_centipede** ----------------L6V2CIDMo&#e-------

- *The originals are HS, IDM, and 6V2IDMo&C#eL; note superimposed cyclic permutation in centipede.*

Deeper comparisons

- **There is a substantial commonality between the vertebrate and arthropod templates**
  - P2V6LmIDMo&BFC#eSAfKbagYpEQqHSLrsJzTR
  - pBEFSJPrHqQTRszSmL6V2IDMo&C#eLfKAbagY

- **The common sequence *bagYp* must antedate the arthropod/vertebrate divergence, and so be approximately 700 million years old.**

- **We can get a clearer view of this relationship by ignoring the tRNAs**

- **72 Artemia_shrimp**   **....r.Q..m.2...**
- **73 notacanthus_shrimp**   **....r.Q..m.2...**
- **74 chrysomya_blowfly**   **....r.Q..m.2...**
- **75 pantodon_butterflyfish**   **..............**
- **76 crioceris_beetle**   **....r.Q..m.2...**
- **77 tribolium_flour_beetle**   **....r.Q..m.2...**
- **79 bombyx_silkworm**   **....r.Q..m.2...**
- **80 Caelorinchus_fish**   **..............**

- **The modified sequences are Qqr --> rqQ and 26m --> m62, and hence *reversals*. The appearance of cyclic permutations and reversals calls for a bit of theoretical reflection.**
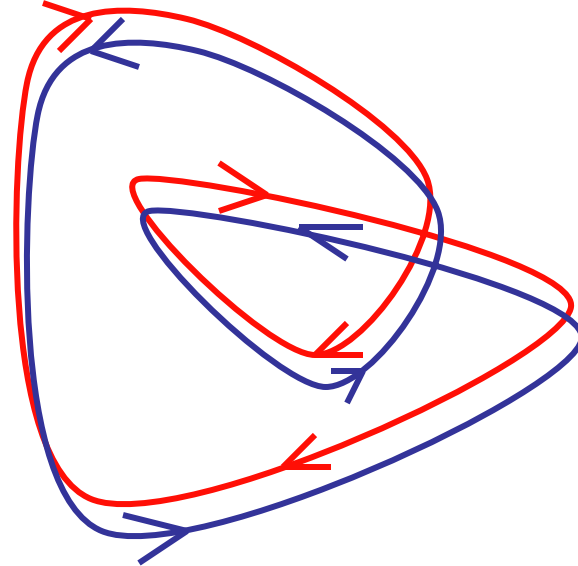
Some theoretical speculations and predictions

- **The mitochondrial DNA strand is circular, double stranded, and oriented**
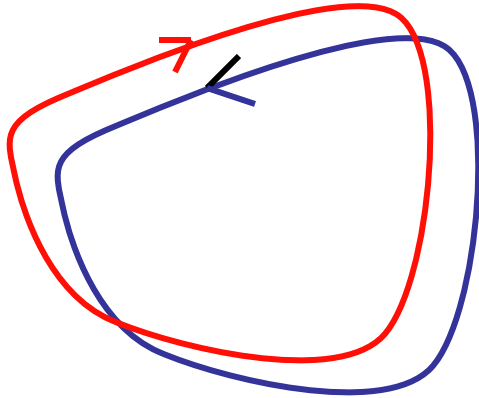

It can fold over on itself.

Re-association is then possible at the point of self-intersection, if orientation is preserved, by switching strands if necessary.

- **Two equiprobable reassociations (Holliday crossovers) are possible, one of which reverses the orientation of the folded-over section, while the other separates the folded-over section as a circle.**

Two crossovers are posited

- **The first crossover changes the original red-strand sequence ABC to *CBA*,**
  - also reversing strand.

- **After rejoining at another point, the second crossover causes a circular permutation of a sequence like ABCD to DABC or CDAB -- without reversing strand**
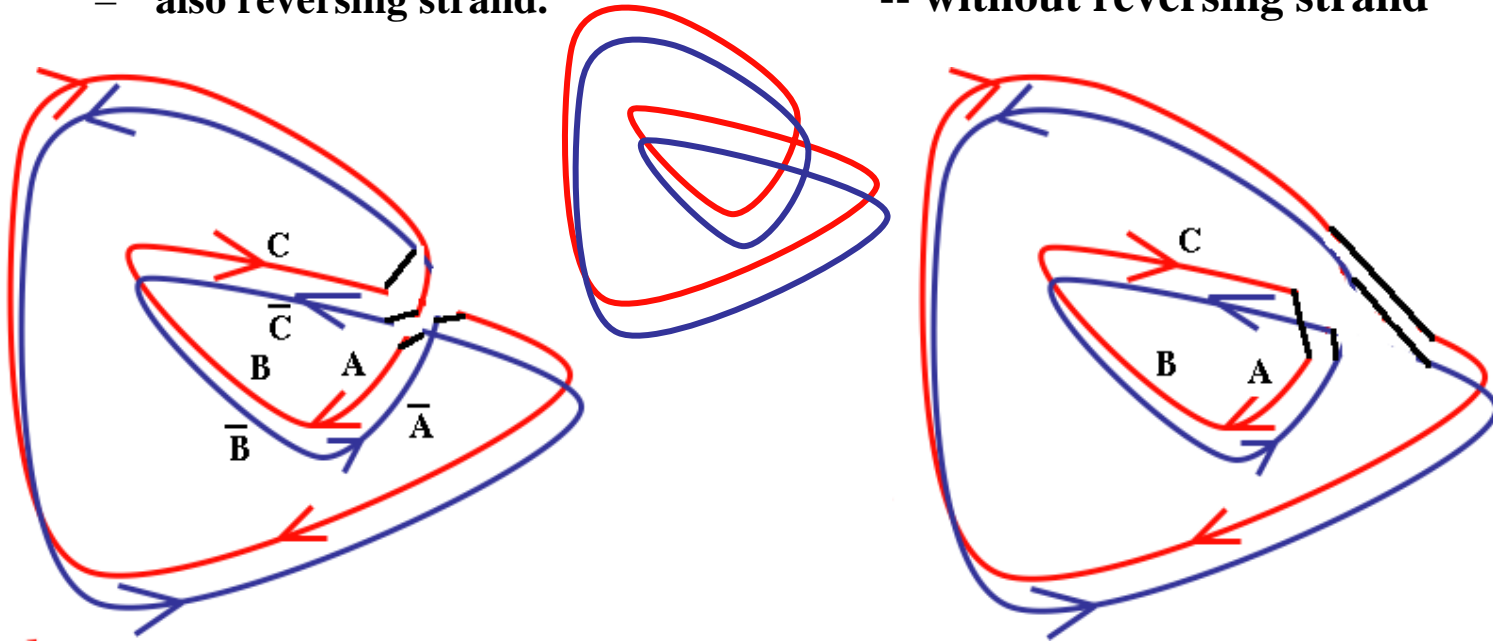
- **Crossovers within genes will destroy them and be fatal, but crossovers at gene ends are OK.**

- **Predictions for DNA:** *two operations on the DNA sequence will be common:*
(1) circular shift of a subsequence (without change of DNA strand) and (2) reversal of a subsequence, with gene motion to the complementary strand.

Finer comparisons: including strand reversal data

- **Additional observations:**
  - Single-gene strand reversal is possible as a degenerate case
  - In a reversing pair, either both or neither gene should change strands

- **Informal statistics for 365 mitochondrial genomes (omitting plants, fungi, protozoa, various cases in which data is suspect)**
  - 25 bad or out-of-range data cases
  - 19 puzzling cases
  - 284 - 44 = 240 successfully analyzed cases
  - 45 circular permutations
  - 7 reversals of more than 1 letter (including 3 pairs)
  - 45 1-letter strand reversals (including one alarming case, Aulopus_fish, with 8 of them)
- **The most convincing gene order reversals associated with strand reversal are:**
  - Asterina among urchins:
  - RDFLB&CVMA#YLmIo6 --> 6oImLY#AMVC&BLFDR
  - The genes which change strand are RD LB&CVMA#Y mIo6 Pg (3 discrepancies)

  - The vertebrate-to-arthropod transition (next slide)

Closer examination of the vertebrate-to-arthropod transition

- The reversals **Qqr --> rqQ** and **26m --> m62** are seen in the no-tRNAs view.
- In the with-tRNAs view, these are **2V6Lm --> mL6V2** and **QqHSLr --> rHqQ**

- The following sequence of steps could account for this transition:

  - **fKbagYpEQqHSLrsJzTRP2V6LmIDMo&BFC#eSA**   mL6V2 reverses
  - **fKbagYpEQqHSLrsJzTRPmL6V2IDMo&BFC#eSA**   BFC#eSAfKbagYp rotates 2
  - **bagYp, mL6V2IDMo&, TR, and fK now remain undecomposed**
  - **fKbagYpBFEQqHSLrsJzTRPmL6V2IDMo&C#eSA**   fKA rotates 1
  - **fKAbagYpBFEQqHSLrsJzTRPmL6V2IDMo&C#eS**   SfKAbagYpBFE rotates 1
  - **fKAbagYpBFESQqHSLrsJzTRPmL6V2IDMo&C#e**   SLrsJzTRP rotates 2
  - **fKAbagYpBFESQqHrsJzTRPSLmL6V2IDMo&C#e**   QqHr reverses
  - **fKAbagYpBFESrHqQsJzTRPSLmL6V2IDMo&C#e**   LmL6V2IDMo&C#e rotates 1
  - **fKAbagYpBFESrHqQsJzTRPSmL6V2IDMo&C#eL**   JzTR rotates 1
  - **fKAbagYpBFESrHqQszTRJPSmL6V2IDMo&C#eL**   rHqQszTRJP rotates 2
  - **fKAbagYpBFESJPrHqQszTRSmL6V2IDMo&C#eL**   szTR rotates 2 (step 10)
  - **fKAbagYpBFESJPrHqQTRszSmL6V2IDMo&C#eL**

- What is the evidential value of such a conjecture?

Estimating the evidential value of a conjectured sequence

- **After its second step, fK, bagYp, TR, mL6V2IDMo&, and BF remain undecomposed, so 21 items are being permuted, but in a restricted way, namely every permutation is a rotation or reversal of an interval. Thus each permutation chosen is one of 21 * 20 * 2 possibilities.**

- **After 8 steps this will generate at most 820^8 possibilities, out of the 37! possible permutations. So the chance of reaching the desired endpoint is**

    - **$\exp(8 \log(820) - 37 (\log 37 - 1)) = \exp(-78.8)$**

- **Thus the two first steps conjectured are very special, and so have evidential weight. A similar argument could be used to support the second reversal.**

- **The genes found to have reversed their strand are mL6V2 QqHr BFJPSs, in weak agreement with what is predicted.**

Befuddlements

- **Why are the lice and thrips so different from the other insects?**

  - Barklouse: Pterygota; Neoptera; Paraneoptera; Psocoptera; Trogiomorpha; Lepidopsocidae.
  - Wallaby Louse: Pterygota; Neoptera; Paraneoptera; Phthiraptera; Amblycera; Boopidae; Heterodoxus.
  - Head louse: Pterygota; Neoptera; Paraneoptera; Phthiraptera; Anoplura; Pediculidae; Pediculus.

- **34 heterodoxus_louse**   **DVsRTeCM&JEHgSmzLqQoYLrKFB#fpAIPS62ab**
- **231 lepidosocid_barklouse   DoC#eLKAbagESJSIM&fYpBFPrHqQTRszmL6V2**

- **"To date there are more than 3000 known species of lice and yet many more remain undescribed. With the possible exception of those species that impinge on the activity of humans and their livestock, the true biology of this cryptic group of insects remains obscure...The phylogenetic relationships and classification of the four main groups of lice have been matters of contention for some time…"**

Befuddlements

- **Why are there so many single-gene strand reversals and so few multiple-gene reversals?**
  - **7 reversals of more than 1 letter (including 3 pairs)**
  - **45 1-letter strand reversals (including one alarming case,** Aulopus_fish, with 8 of them)

- **What is one to make of, and how is one to treat, the remaining mysterious cases?**

  - 14 strongyloides_nematode   mB2MCzLfJLP6sYT.QKSrq&ESgaHDF#RoVIpA
  - 17 Laqueus_lampshell   Fe.zHaD&rBsIpTEPJKSQYgAbSofL62MLRm#qC
  - 23 Terebratalia_brachiopod   eCzYagHaD&rLRA2MBsL6JFmV#qPQfIpTEKSo
  - 25 Schistosoma_fluke   RVM&SeT6CC2fs#LSFIPaoBLErYgJ.zQqDKpAm
  - 28 roboastra_gastropod   •••PfYHCDLbFaEJ2MpSSqTgIoKeV6LBRsrm#&
  - 33 Tigriopus_crustacean   eMRAJSCBL.DPmfL#zHEqFg6.T2KIrpVsQbaS&
  - 34 heterodoxus_louse   DVsRTeCM&JEHgSmzLqQoYLrKFB#fpAIPS62ab
  - 35 rhipicephalus_dog_tick   ------mL6V2IDPrHqQTRszSLCMo&#e-------
-    -   38 ciona_seasquirt   £££fmIpVqPE#LSeTrLsMFaofzRQ
  - ....

Befuddlements

- **Can relationships be found between the main template groups?**

  - **Vertebrate**   **P2V6LmIDMo&BFC#eSAfKbagYpEQqHSLrsJzTR**

  - **Urchin**   **P2JTRDFLB&CVMA#YLmIo6eEQfKbagSpqHSrsz**

  - **Gastropod**   **QzACPf#&YHDLbFaEJ2MpSSqTgIoKeV6LRBsrm**

  - **Platyhelminth**   **gHzQqDPMaoVBAmFRIKpS&eT6C2fJs#LSLErY**

  - **Schistosome**   **MaoBAmFRIKp&VSeT6C2fsL#SLErYgJHzQqDP**

  - **Nematode**   **RVsQ&J2SF#maKLSoIEDPzLgTqeCMAYfH6prB**

  - **Roundworm**   **Q2#mPaIYfH6pCSRAVrJSoTqe&sEDzLgKBLFM**

  - **Seasquirt**   **HEDgK6IpJVqPCY#LSerAML2MT&SsYFaoBfzmRQ**

- **For example, Schistosome can be reached from Platyhelminth by the following steps:**

  - **MaoVBAmFRIKpS&eT6C2fJs#LSLErYgHzQqDP**   **VBAmFRIKp rotates**
  - **MaoBAmFRIKpSV&eT6C2fJs#LSLErYgHzQqDP**   **SV& reverses**
  - **MaoBAmFRIKp&VSeT6C2fJs#LSLErYgHzQqDP**   **Js#LSLErYg rotates**
  - **MaoBAmFRIKp&VSeT6C2fs#LSLErYgJHzQqDP**   **#L interchanges**