Blastocystis comparative genomics and evolution

Workshop session 7

Andrew J. Roger, Dalhousie University
Email: andrew.roger@dal.ca

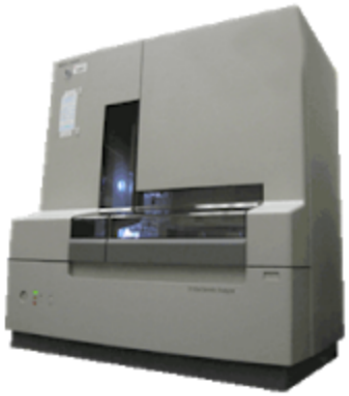
Website: rogerlab.biochem.dal.ca

Go to "Teaching" and follow links to workshop PDF

Blastocystis genomics so far

Technologies have been changing:

- 2011 – *Blastocystis* Subtype 7 (Singapore B)
  - Denoeud *et al. Genome Biol*. 12: R29
  - Sanger sequencing (genome, cDNAs)
- 2015 – Blastocystis Subtype 4 (WR1) genome:
  - Wawrzyniak et al. *Genome Data* 4:22-3
  - Illumina DNA sequencing (Illumina Hi-Seq)
- 2017 – Blastoystis Subtype 1 (NandII) genome:
  - Gentekaki et al. *PLoS Biol.* 15:e2003769
  - Illumina Hi-Seq for DNA and RNA-Seq
- 2018 + Long-read sequencing (10,000 bp+)
  - Oxford Nanopore (MinION) and PacBio



ABI-3130-xl

Illumina HiSeq 2500

MinION

Blastocystis genomes in GenBank

Annotated Genomes

- *Blastocystis* sp. ST7- Singapore isolate B (genome contigsand predicted genes)
- *Blastocystis* sp. ST4-WR1 isolate (genome contigs and predicted genes)
- *Blastocystis* sp. ST1- NandII isolate (genome contigs and predicted genes)

Draft genome assemblies (with no predicted genes):

- *Blastocystis* sp. ST2 (Flemming isolate)
- *Blastocystis* sp. ST3 (ZGR isolate)
- *Blastocystis* sp. ST6 (SSI:754 isolate)
- *Blastocystis* sp. ST8 (Dmp/08-128 isolate)
- *Blastocystis* sp. ST9 (F5323 isolate)

Genome statistics for *Blastocystis* subtypes 1, 4 and 7

Subtype	ST1	ST4	ST7
Genome assembly size	16.5 Mb	12.9 Mb	18.8 Mb
Scaffolds	580	1301	54
G+C content	54.6%	39.6%	45.2%
Number of protein coding genes	6544	5713	6020
Genes with introns	94.6%	92.7%	84.6%
Average exons per gene	6.45	5.06	4.58
Most frequent length of introns (nt)	30 (54%)	30 (36%)	30 (21%)
Number of introns	35,412	24,093	18,200
Number of subtype unique genes	611	221	670

What use are genomes?

- **Learn about the potential biochemical, structural etc. diversity of *Blastocystis* – both *within* and *between* subtypes**
  - What genes are unique to each genome?
  - How does it affect their biology?

- **Shared 'core' genes by all *Blastocystis* but not in other stramenopiles**
  - Probably important for adaptation to gut

- **Identify proteins potentially involved in colonization of the gut:**
  - Nutrient acquisition (carbs, amino acids, metabolites, transporters)
  - Immune system evasion
  - Virulence? (could differ a lot between strains and subtypes)
  - Enzymes of anaerobic energy metabolism and oxygen detoxification
  - Proteins/metabolites that could affect other gut microbes (e.g. polyketides)

- **Understand the mechanisms by which new 'types' of *Blastocystis* emerge**
  - Gene duplication, loss and gene acquisition by horizontal (lateral) gene transfer
  - Recombination?

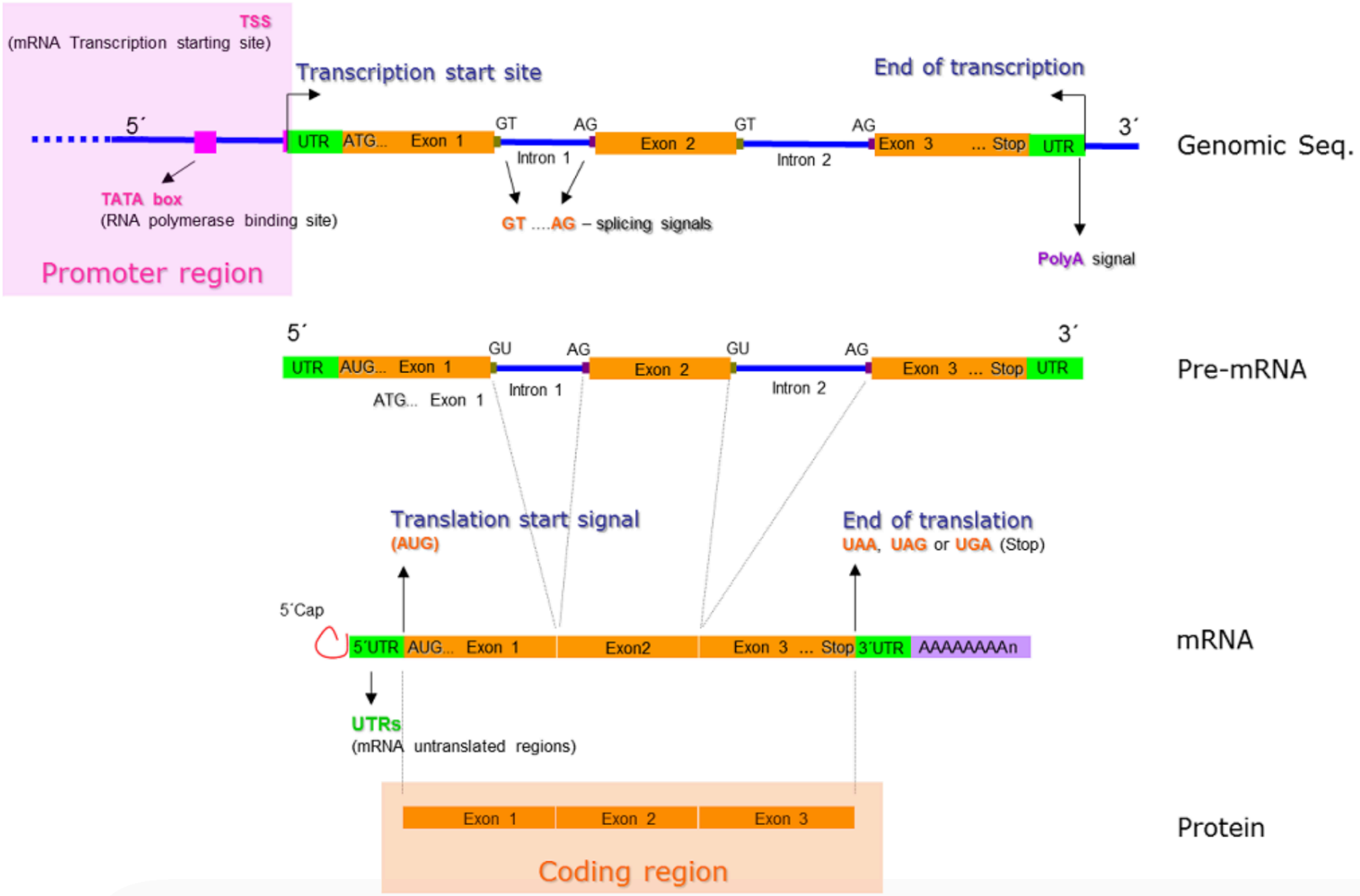
- **Reference genomes for metagenomic investigations of role in microbiome**

Overview

All parts are contained within the PDF with hyperlinks to online tools, databases and files you need:

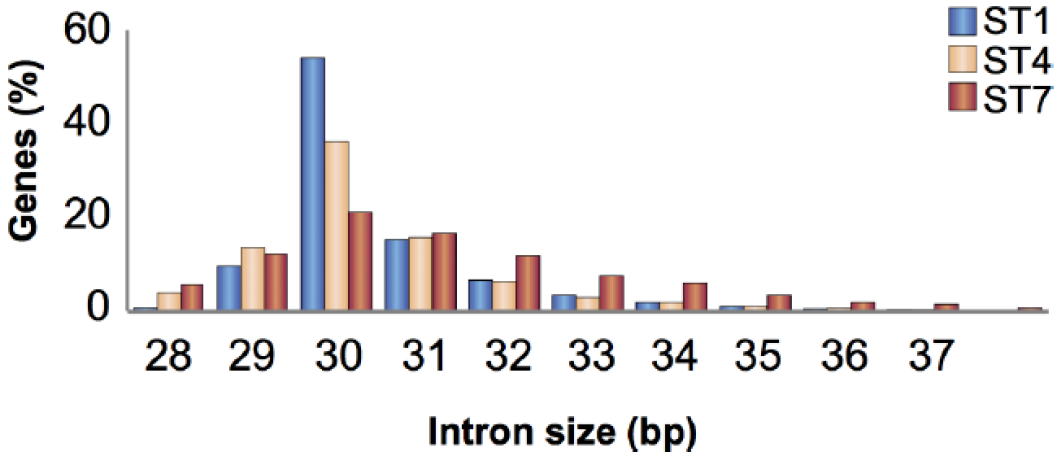
- Part 1 – browsing GenBank and genome sequences
  - NSCBI Genomes, Artemis browser
- Part 2 – predicting protein function
  - BLAST, Interproscan, eggNOG-mapper
- Part 3 – predicting protein subcellular localization
  - TargetP, TMHMM, BUSCA
- Part 4 – multiple alignment & phylogenetics of protein family
  - MAFFT, distance methods (NJ) and IQ-TREE (ML)
- Part 5 – 'typing' carbohydrate active enzymes
  - CAZy, and dbCAN2
- Part 6 – 'typing' peptidases/proteases
  - MEROPS database

Eukaryotic genome/gene structure



Interesting *Blastocystis* genome features

Small introns



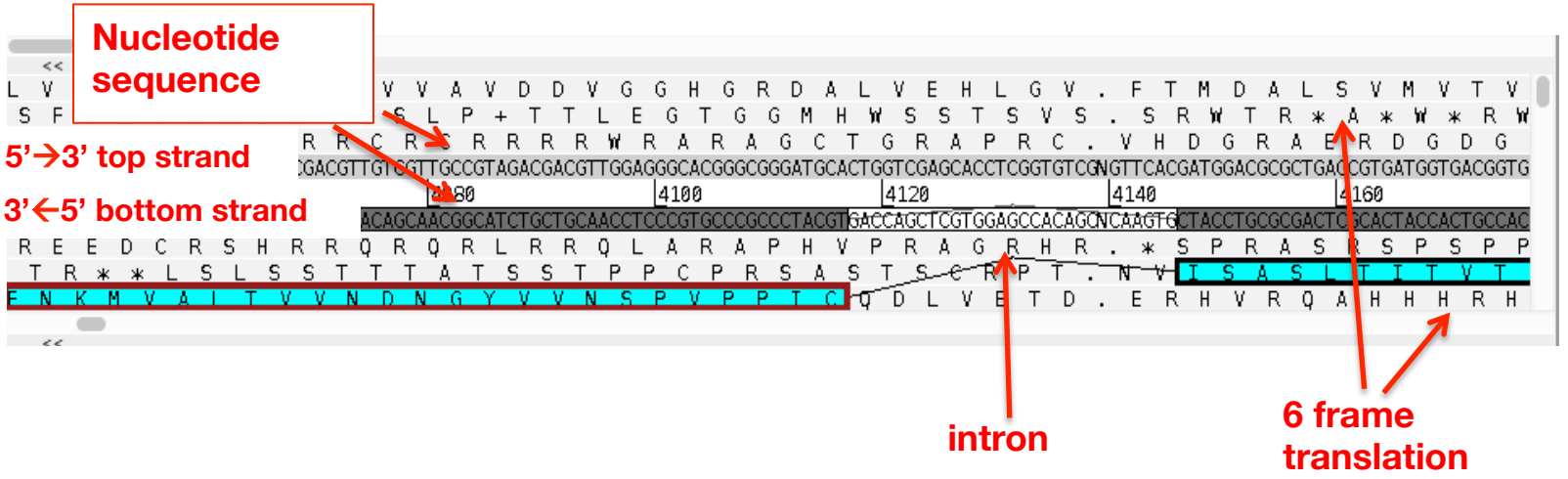
Stop codons in 15-25% of genes are created by polyadenylation

TGTTTGTT motif 5 nucleotides downstream of polyadenylation site

Genomic sequence →
mRNA sequence →
Amino acid sequence →

Reading frames

Protein-coding are translated from the mRNA in triplet nucleotide sequences (codons). There are 6 different 'reading frames' -- only one is 'correct'



Because intron lengths are often NOT multiples of 3, the correct reading frame will often change between exons and introns

Note: gene models (start codons, exons/introns and stop codons) are bioinformatic PREDICTIONS and can be wrong (you can edit them in Artemis)

FASTA format

Description line (starts with '>')

```
>AV274_0014 AV274_0014 homeobox protein TGIF2
MNDSNQSGRSESTIEMEIYQDDIQNCIALSVDSRIEKMVELTSDIEELLGLQSDKEFRMQ
RVNLFRQRAVAEGIPPVLDPEFSVKVENYCSLLQSKKAILLSMYKCCEDFCLAMHNELEA
INQSFANNPEERAAFVDNYMRSSCSRSRCSKLASGKHQRRNRLPSHALSILWDFVRTHKK
NPYPSTQQKEALARQTNLTMTQIRNWFTNTRKRKLSQAPESDEDYSIGSDNEDSYPSPPP
EESAGRRSLKRRAVGRNAKTAKARKQSVDLTPVLPLPFAPNAVPDDAPQNRTANASHASH
AGRWIQHAEGGMLSSVDGGKSSEKMEEAEAPLEMKTESASHVGTPRFDRDFSLFEPGSIP
NSGIGFSLSHFSDDPLLLGLSLGDLQDNEFFNNNPIAMRKKDGEEGEAIPPHLDEERVDM
NSDTIVPSSV
```

amino acid sequence

Scoring sequence similarity

BLAST seq. scoring approach

Score	Expect	Method	Identities	Positives	Gaps
244 bits(624)	6e-71	Compositional matrix adjust.	132/300(44%)	189/300(63%)	16/300(5%)

```
Query  17   EIYQDDIQNCIALSVDSRIEKMVELTSDIEELLGLQSDKEFRMQRVNLFRQRAVAEGIPP  76
            E+ Q D++NC++L++DS   ++VELTSDIEE LGL SD+E+RM R+++ R+ A+ + I P
Sbjct  238  ELPQADVKNCLSLALDSHASRIVELTSDIEEQLGLVSDREYRMSRLSVLRKSALDQCIFP  297
```

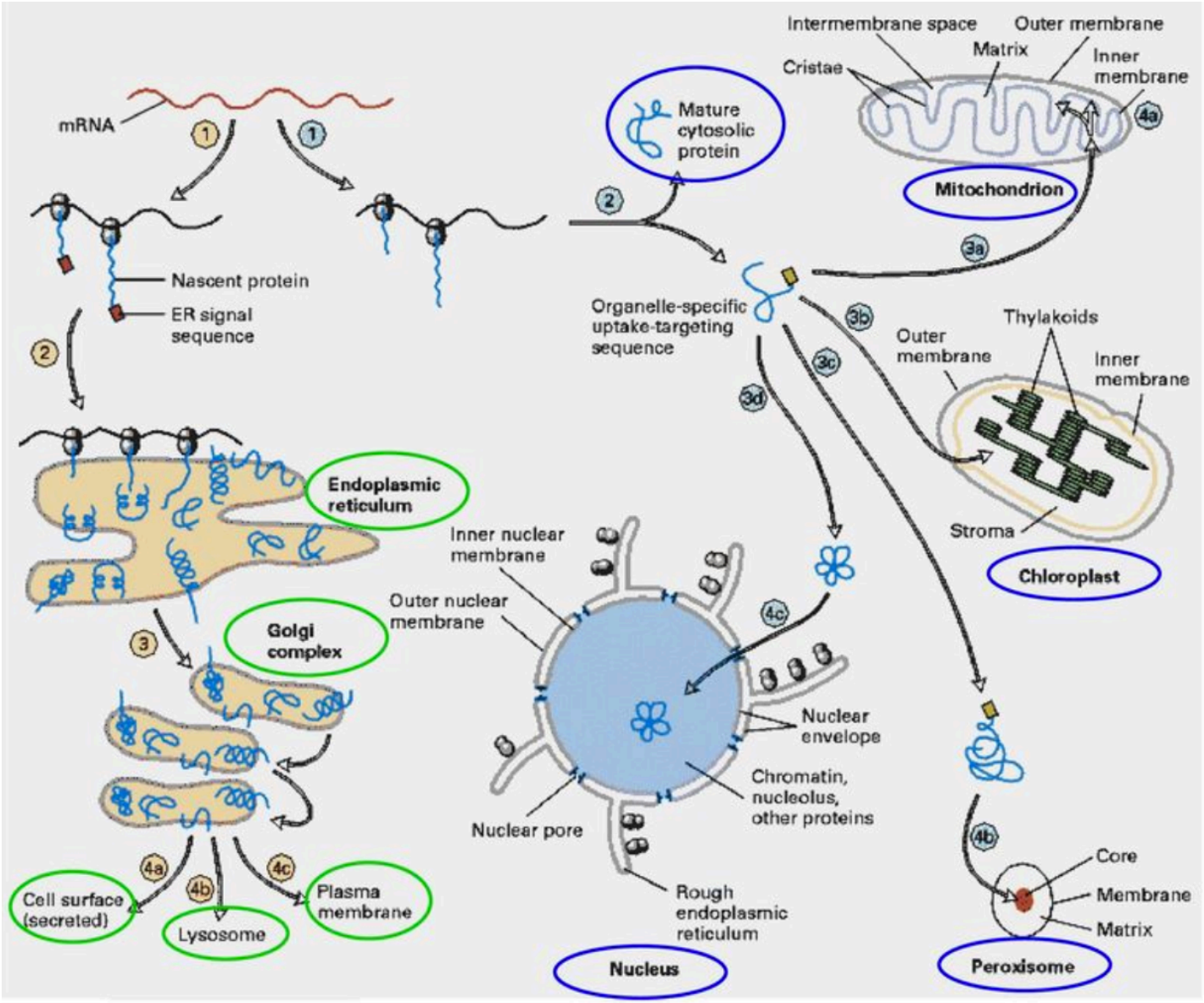
Similarity scoring is based on a general 'scoring matrix' (BLOSUM62) that upweights common amino acid interchanges between amino acids and downweights uncommon amino acid interchanges

---

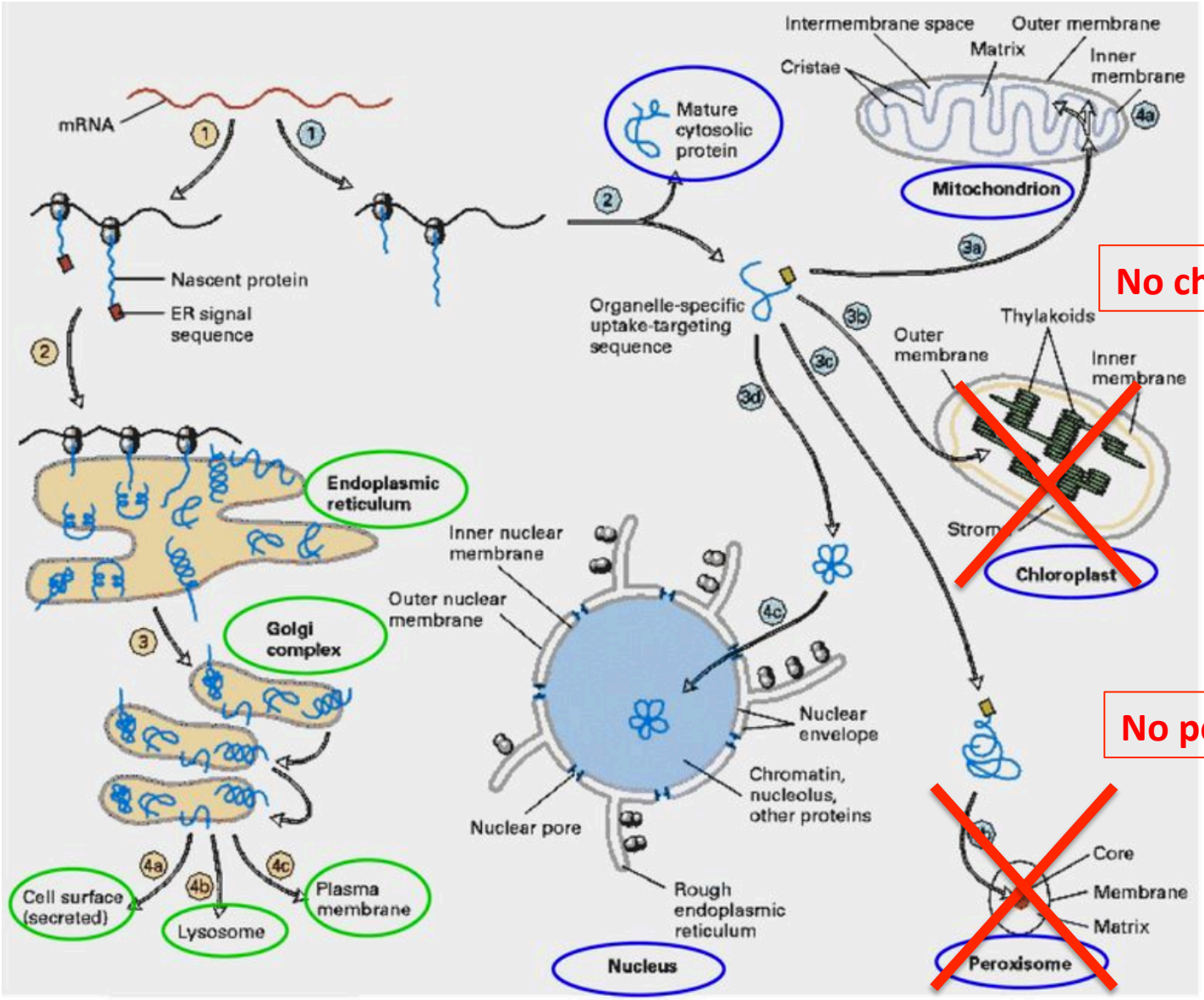
Hidden Markov Model (HMM) (statistical model of a multiple alignment)

Similarity scoring is based on how query sequence matches the conservation of particular amino acids in the multiple alignment. Every position in the alignment has a separate 'scoring' system based on frequencies of amino acids at that site

Protein localization/targeting



Protein localization/targeting in Blasto



Subcellular localization/targeting

- ER signal peptide (N-terminus)
  - 5–30 mostly hydrophobic amino acids forming a helix, often preceded by one or more basic amino acids
  - direct proteins into ER cotranslationally, then Golgi and endomembranes or secretion to plasma membrane or outside of cell
  - Cleaved
- Mitochondrial targeting peptide (N-terminus)
  - 10-70 Amphipathic helix (alternating hydrophobic amino acids and positively charged (R, K) and sometimes hydroxylated (S,T)
  - Cleaved
- Transmembrane helix
  - 20-25 hydrophobic amino acids forming a helix that spans a lipid bilayer

Orthology and paralogy

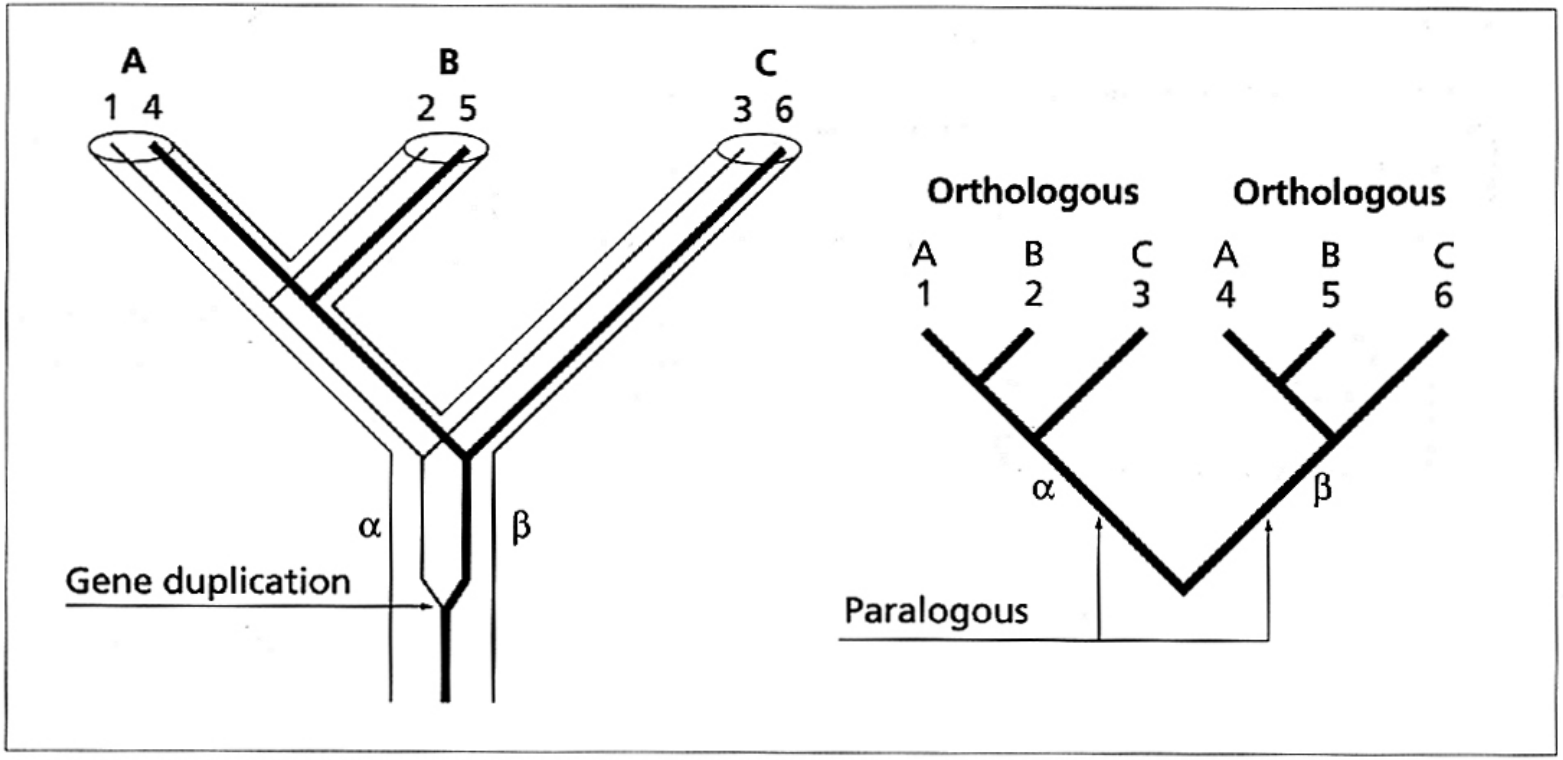


Fig. 2.22 Phylogeny for three species A–C and six genes that stem from a gene duplication resulting in two paralogous clades of genes, α and β. The α genes 1–3 are orthologous with each other, as are the β genes 4–6; however each α gene is paralogous with each β gene as they are separated by a gene duplication event, not a speciation event.