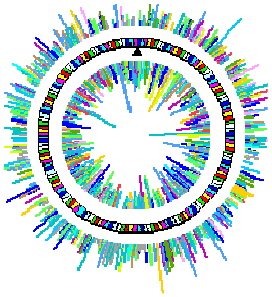
Comparative Gene Finding in Yeast

Marina Alexandersson

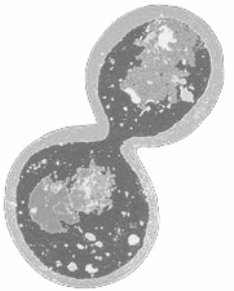
Oslo, June 2, 2005

2005-06-02

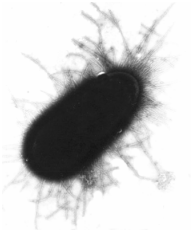


Sequencing history

1995: Two microbial genomes (1.8, 0.6)

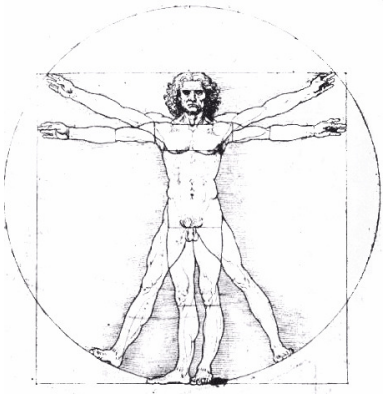


1996: Saccharomyces cerevisiae (12)

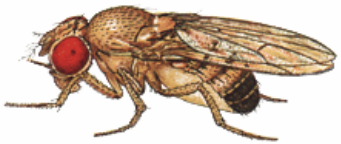


1997: E. coli (4.6)

1998: C. elegans (97)



2000: Drosophila melanogaster (180)



2003: Homo sapiens (3,200)

ITWM
Fraunhofer
CHALMERS
Research Centre
Industrial Mathematics

More to come...

ITWM
Fraunhofer
CHALMERS
Research Centre
Industrial Mathematics

Whole genome analysis

- Gene finding
- Sequence alignment
- Regulatory region discovery
- Genome evolution

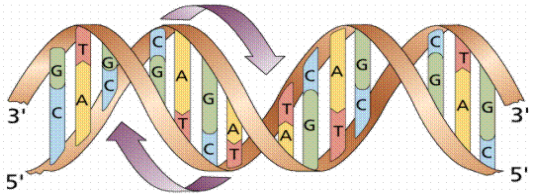
Whole genome analysis - Why?

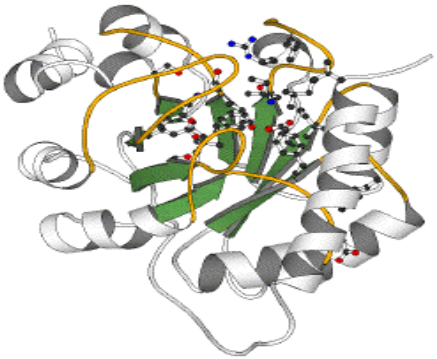
- Complete gene and protein sets
- Primary sequence of all genes
- Sequence relationships between genes and proteins
- Function of new proteins
- Transcriptional level of all genes
- Understanding methabolic pathways
- Trace disease genes
- ...

Gene finding

Fraunhofer CHALMERS
Research Centre
Industrial Mathematics

Gene expression





DNA

transcription

RNA

translation

Protein

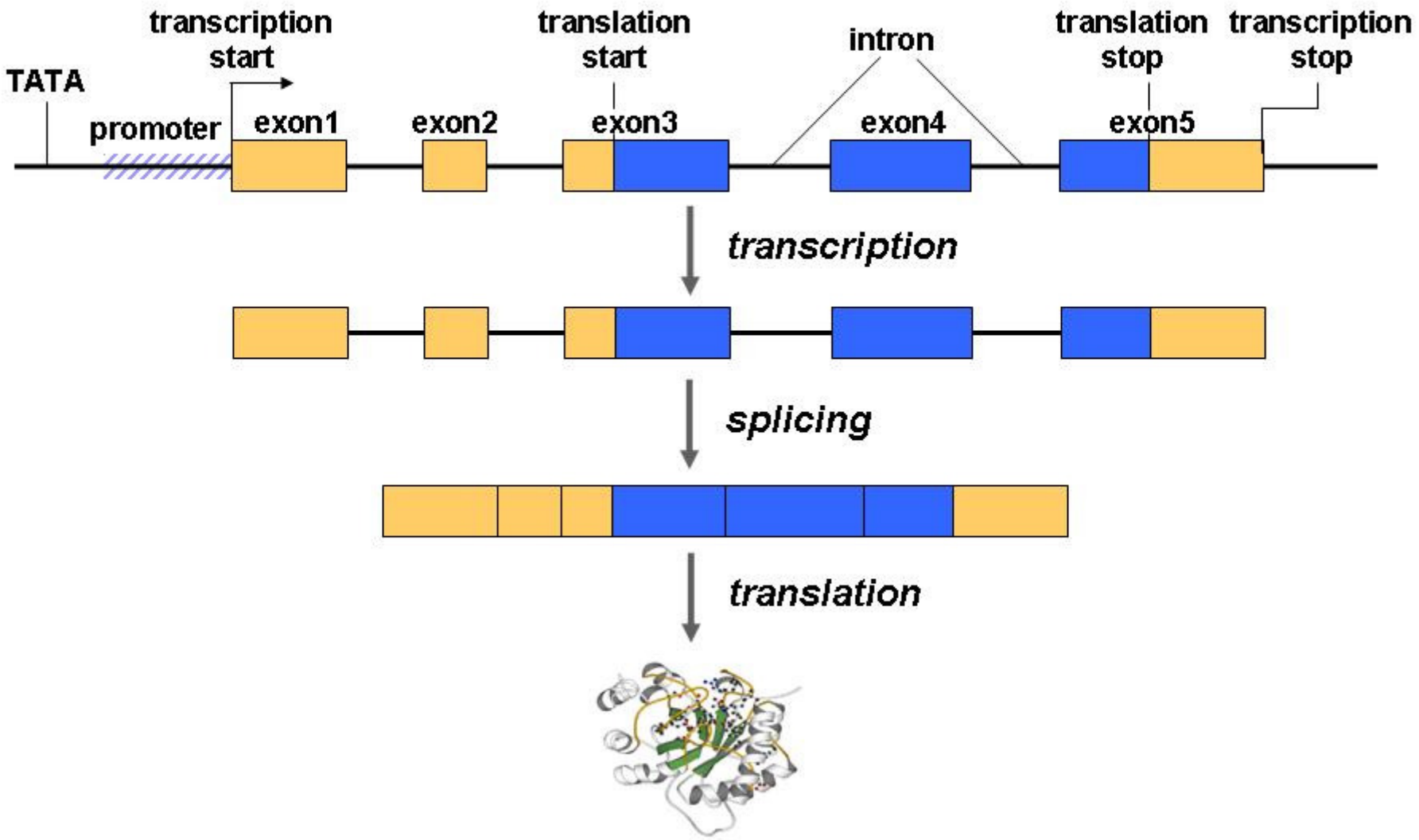
CCTGAGCCAACTATTGATGAA

CCUGAGCCAACUAUUGAUGAA

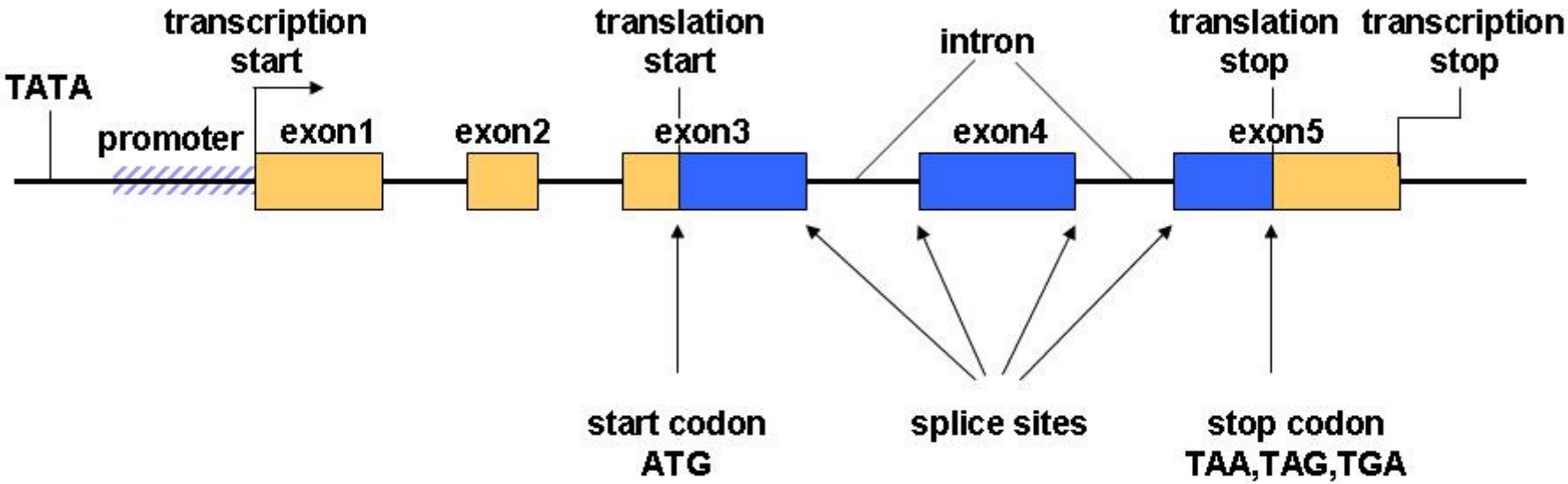
PEPTIDE

Fraunhofer ITWM
CHALMERS
Research Centre
Industrial Mathematics

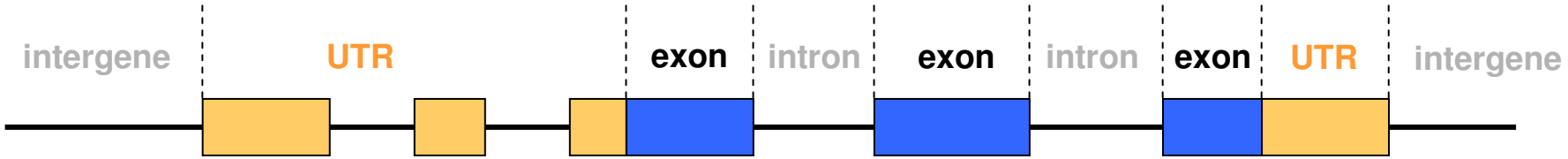
Gene structure (in higher organisms)



Finding genes



Hidden Markov models (HMMs) for gene finding

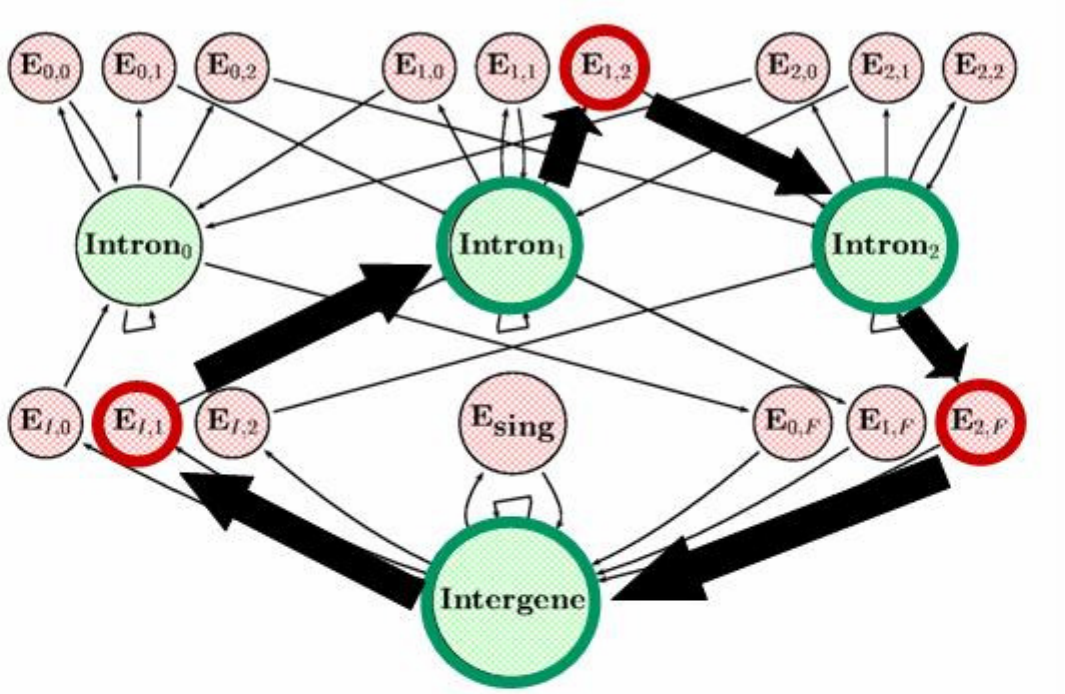


Observed:

CTTGATGCTGGCACGTTCTGCTTCATCGGAGACAAATTACGGCTTTCCGGAGCA

Hidden:

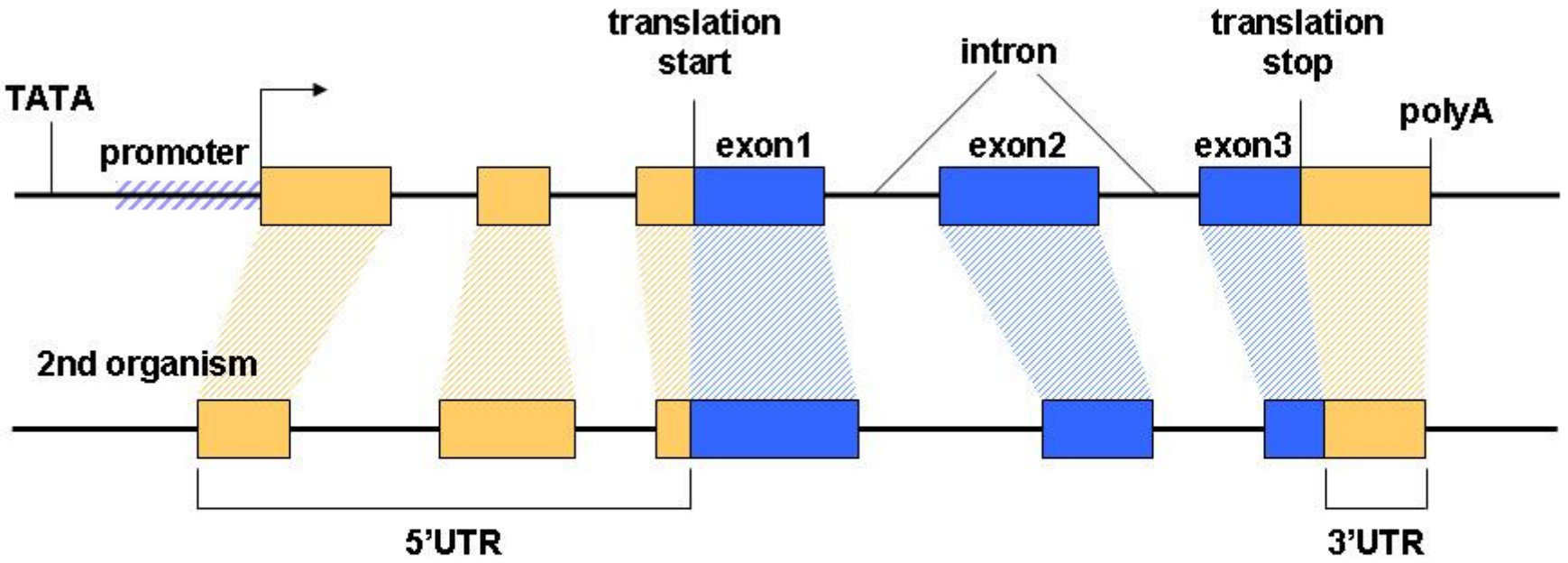
CTTGATGCTGGCACGTTCTGCTTCATCGGAGACAAATTACGGCTTTCCGGAGCA



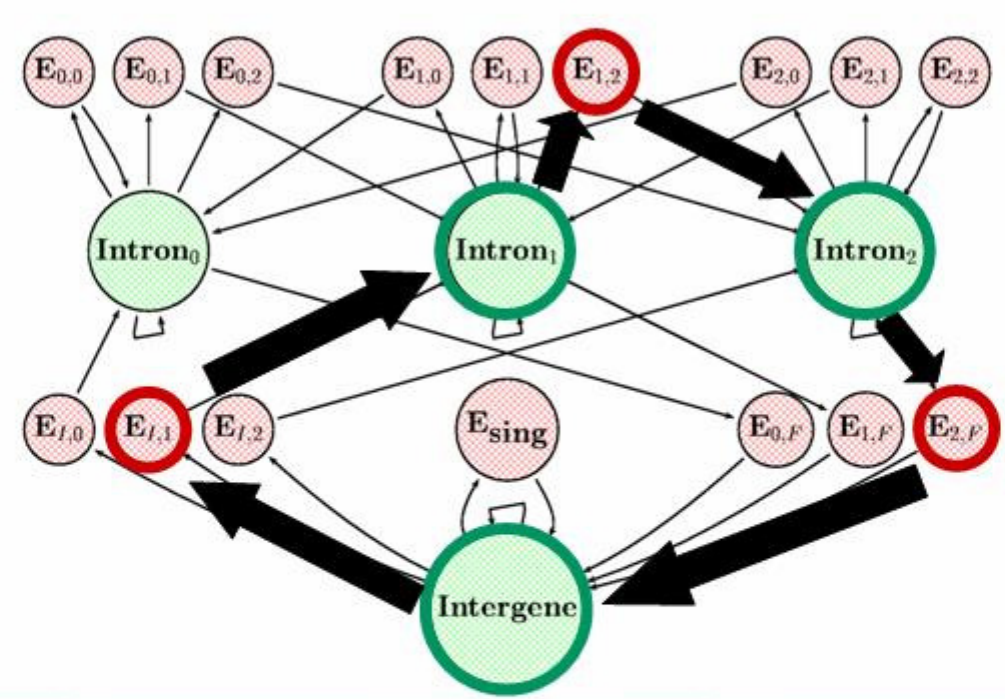
Fraunhofer ITWM

CHALMERS
Research Centre
Industrial Mathematics

Comparative gene finding



Generalized Pair HMMs



Fraunhofer ITWM

CHALMERS
Research Centre
Industrial Mathematics

Gene finding in yeast

Fraunhofer CHALMERS
Research Centre
Industrial Mathematics

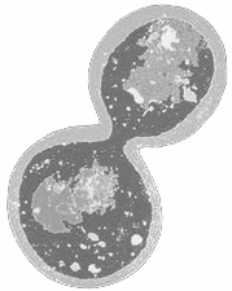
Saccharomyces cerevisae

- The baker's and the brewer's yeast

Yeast genetics

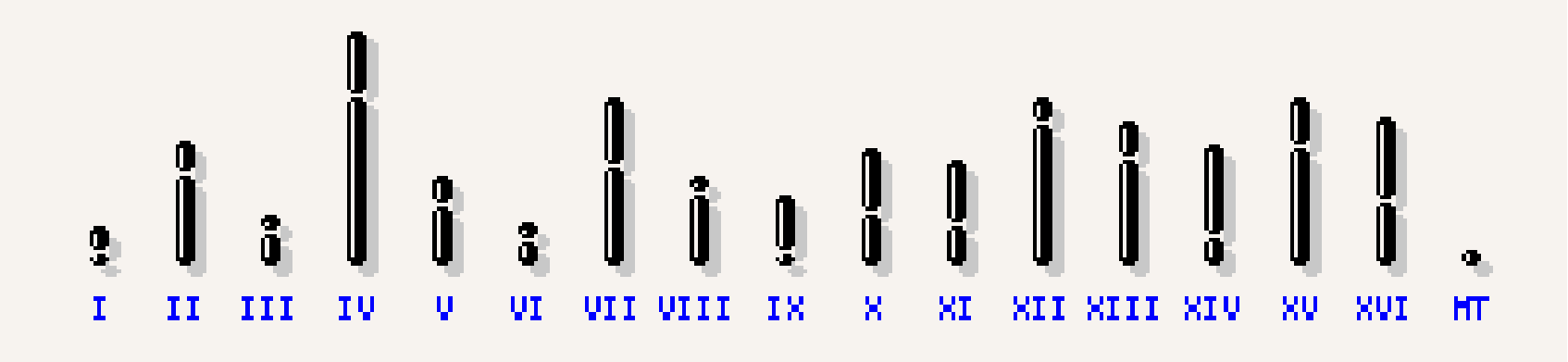
- A yeast is a stage in the life-cycle of fungus where it lives in a single-celled state.
- True yeasts reproduce by budding
- Habitats: on leaves and flowers, soil and salt water, skin surfaces and intestinal tracts

Why yeast as genetic models?



- Basic cellular mechanisms conserved
- Unicellular
- Grow on readily controlled, defined media
- Ideal life cycle
- Very compact genome
- Quick to map a phenotype producing gene
- Single gene deletion mutants
- One third of the genes have counterparts in human

Saccharomyces cerevisiae



Saccharomyces cerevisiae

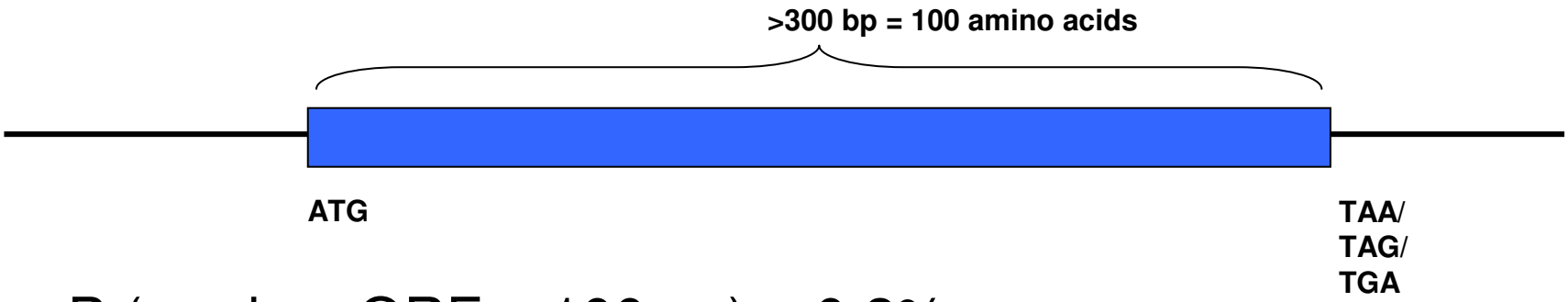
- 16 chromosomes ~ 12 Mb
- 4800 – 6000 protein coding genes
- 70% coding DNA
- 5% of the genes have introns

Human

- 46 chromosomes ~ 3.2 Gb
- 25’000 – 30’000 genes
- 3% coding DNA
- >85% ??

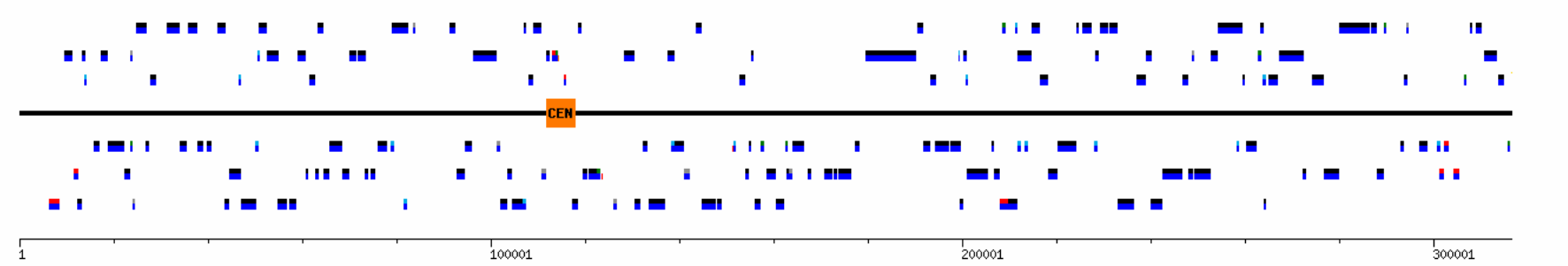
Gene finding in *S. cerevisiae*

- Open Reading Frames (ORFs)



- P (random ORF > 100 aa) < 0.2%

Gene finding in *S. cerevisiae* (cont.)



- 7472 ORFs > 100 amino acids (Goffeau et al. 1996)
- 3000 overlapped
- Adjusting for overlaps: 6275 ORFs
- Using measures of coding capacity: 5885 putative protein-coding ORFs

The mystery of ’orphans’

- Out of the 5885, half were ’orphans’ with unknown function or homology
- The number of orphans grew faster than the number of homologues
- Suggested explanation: true number of protein-coding genes ~ 4800.

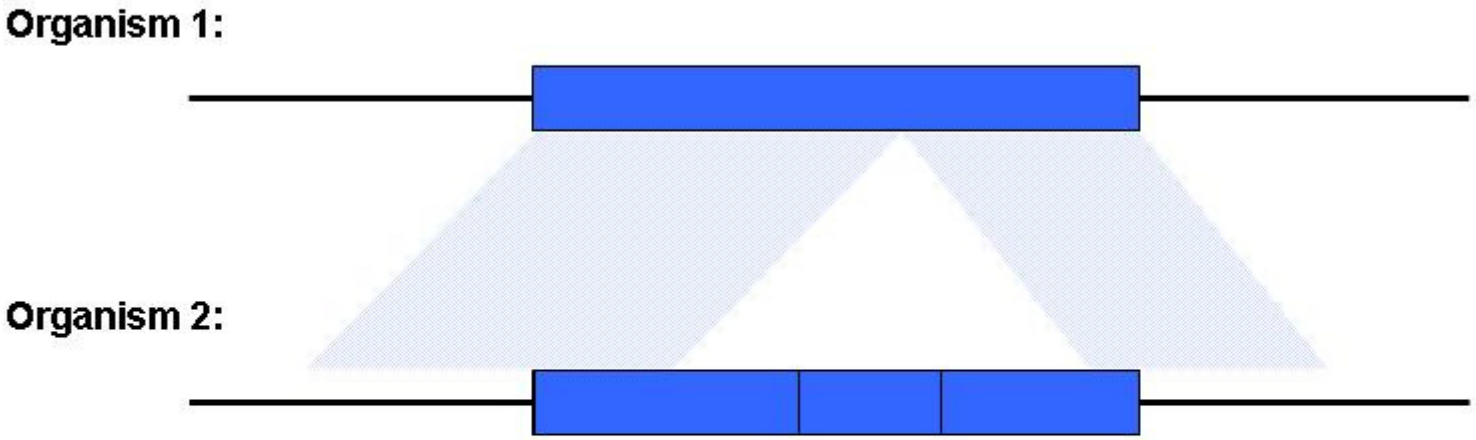
Estimated no. genes in *S. cerevisiae*

- Goeffau et al. (1996) 5885
- Cebrat et al. (1997) ~4800
- Kowalczuk et al. (1999) >4800
- Blandin et al. (2000) 5651
- Zhang & Wang (2000) 5645
- Wood et al. (2001) <5570
- Mackiewicz et al. (2002) 5322
- Kumar et al. (2002) ~6000
- Kellis et al. (2003) 5726

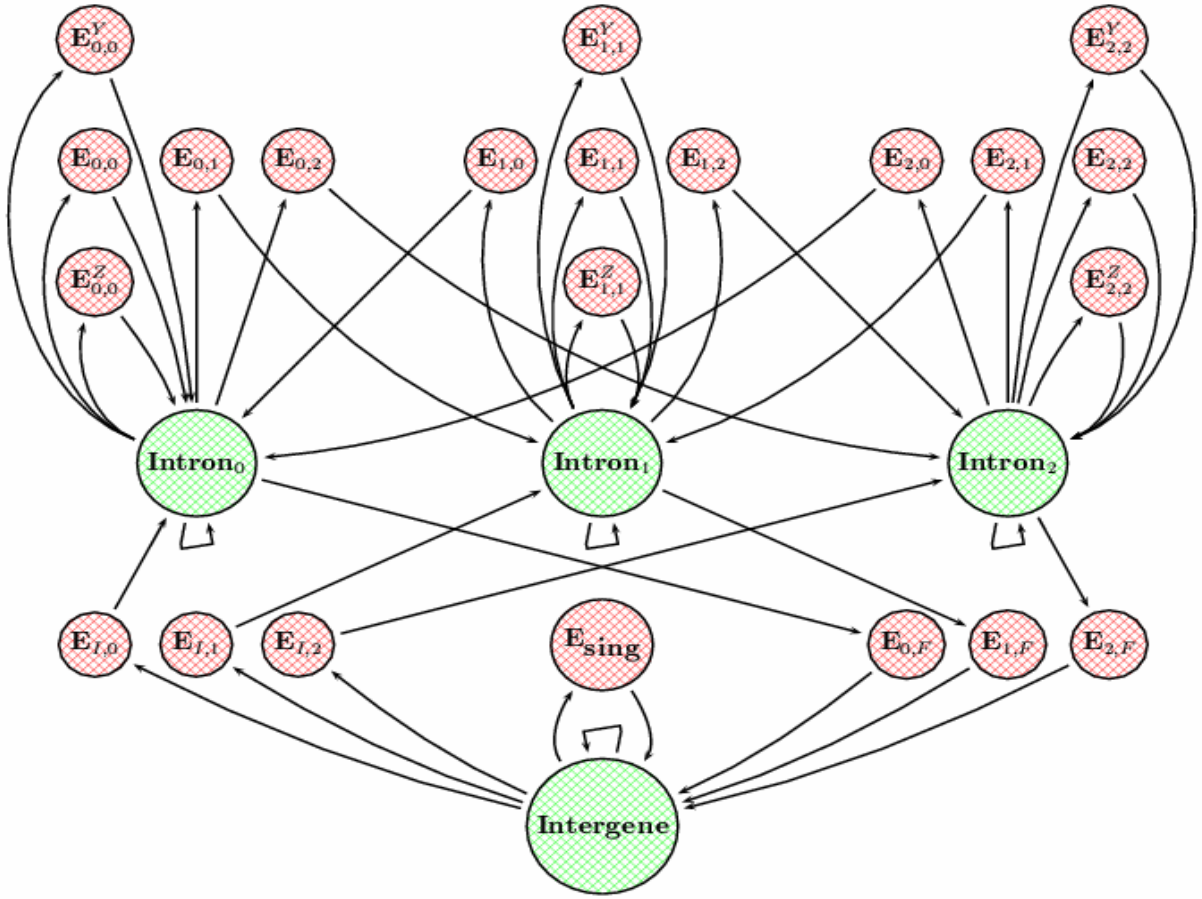
Difficulties in comparative gene finding

Fraunhofer CHALMERS
Research Centre
Industrial Mathematics

Insertions and deletions of exons and introns



Insertion of introns



ITWM

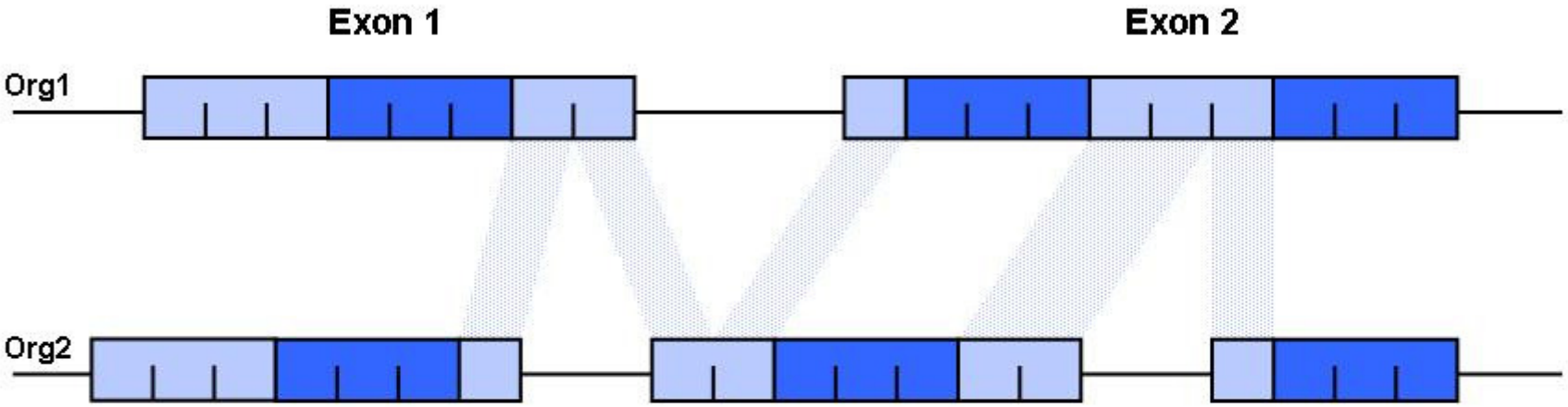
Fraunhofer CHALMERS
Research Centre
Industrial Mathematics

Genome rearrangements

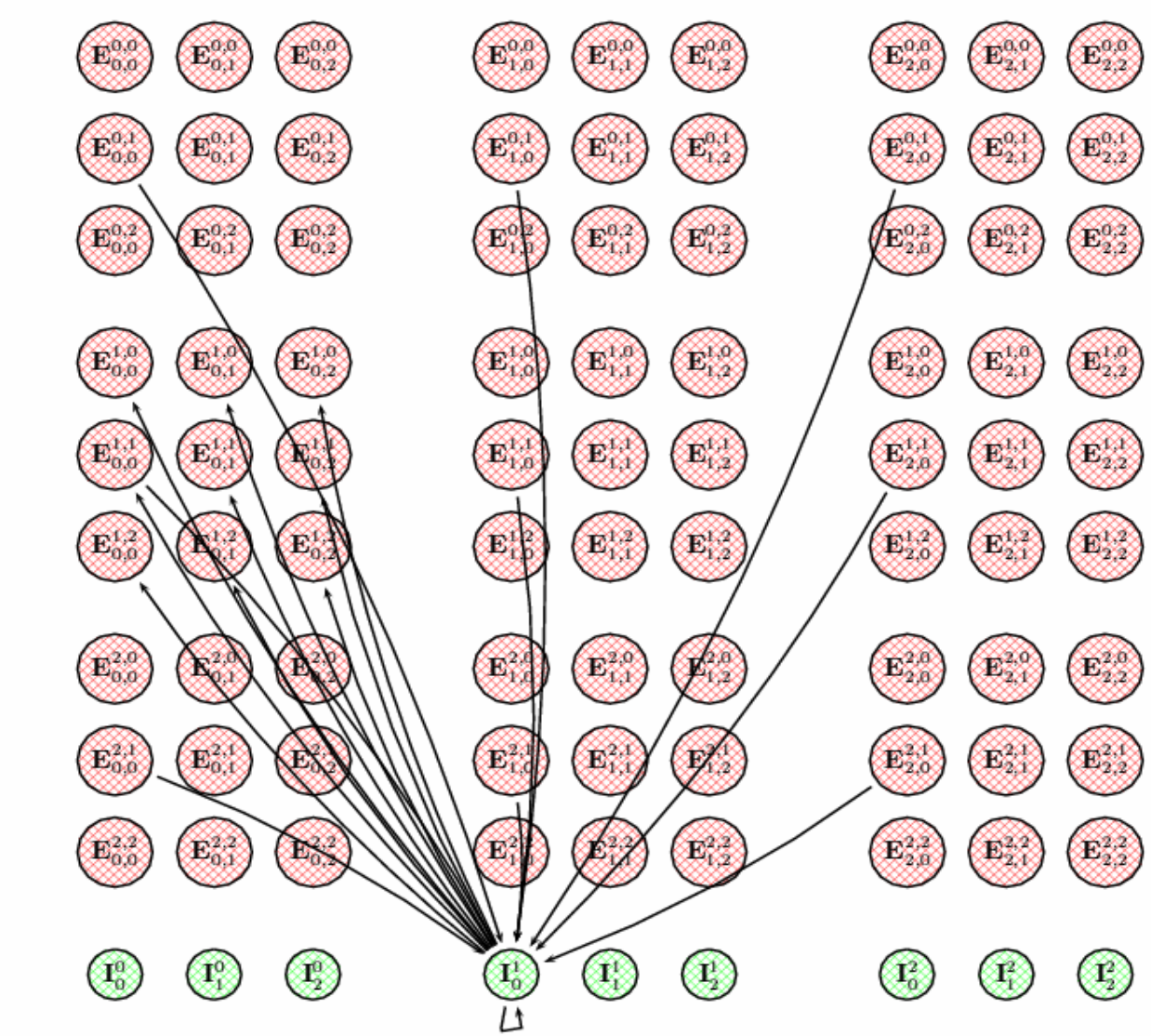
Organism 1:

Organism 2:

Frameshifts



The frameshift model



Comparative gene finding

- De novo – using statistical patterns in the DNA sequence.

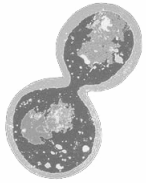
  + **Identifying complete gene structure**
  - **Computationally complex**

- Homology searches – employ sequence similarity between evolutionary related organisms or proteins.

  + **Validating potential genes**
  - **Dependence on database**
  - **Fail to resolve complete gene structure**

- Comparative methods – using multiple sources

  + **Very specific**
  - **Computationally complex**



S. cerevisiae versus *C. elegans*

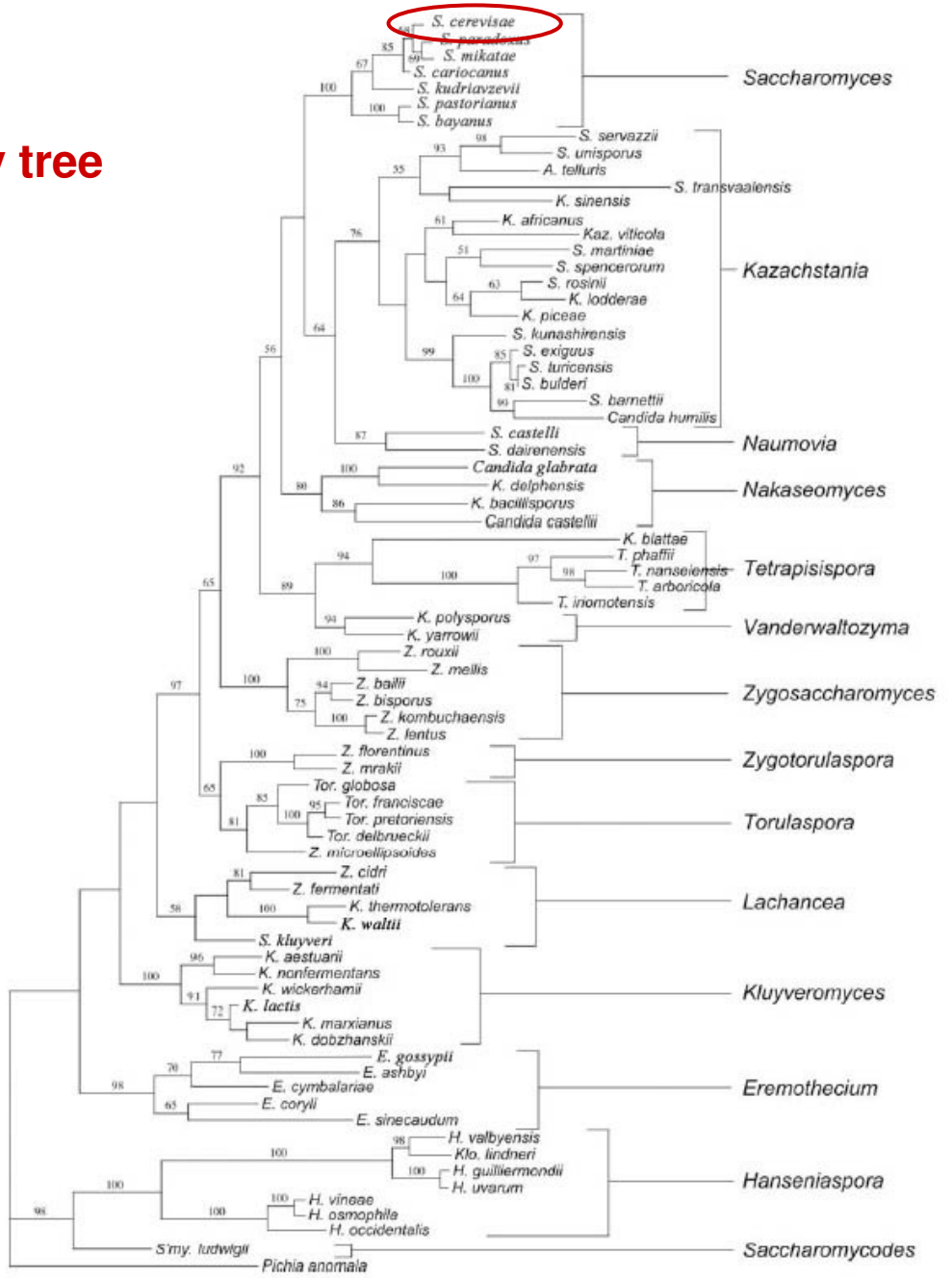
- Substantial fraction of genes with one-one relationships
- Responsible for core biological processes such as
  - various metabolisms
  - protein folding
  - degradation
- Processes carried out by a similar number of proteins.

Comparative analysis in yeast

- Yeast provide excellent model systems for the development of comparative analysis tools:
  - Multiple whole genome comparisons
  - Rearrangements
  - Insertion and deletions of entire exons or introns
  - Regulatory elements
  - Small coding genes (smORFs)

Fungi

- evolutionary tree



Complete fungi genomes

- Saccharomyces
  - Complete: 1 (*S.cerevisiae*)
  - Draft: 5
  - In progress: 2
- Other fungi
  - Complete: 8
  - Draft: 18
  - In progress: 20

Comparative tools

- Fungal BLAST
- ClustalW
- ORF finders
- Generalized Pair HMMs
- Genome Browsers
- ...