

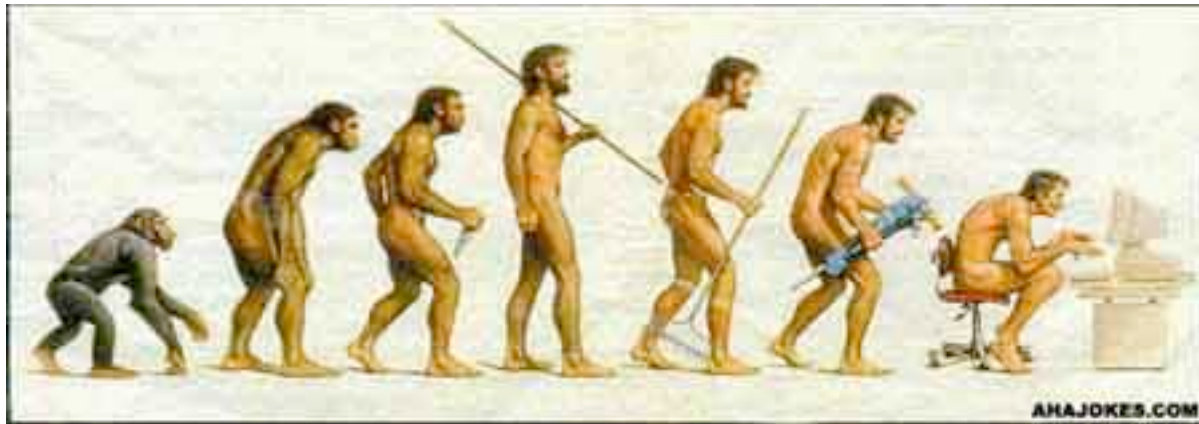
Large scale genomes comparisons Bioinformatics aspects (Introduction)

Fredj Tekaiia
Institut Pasteur
tekaia@pasteur.fr

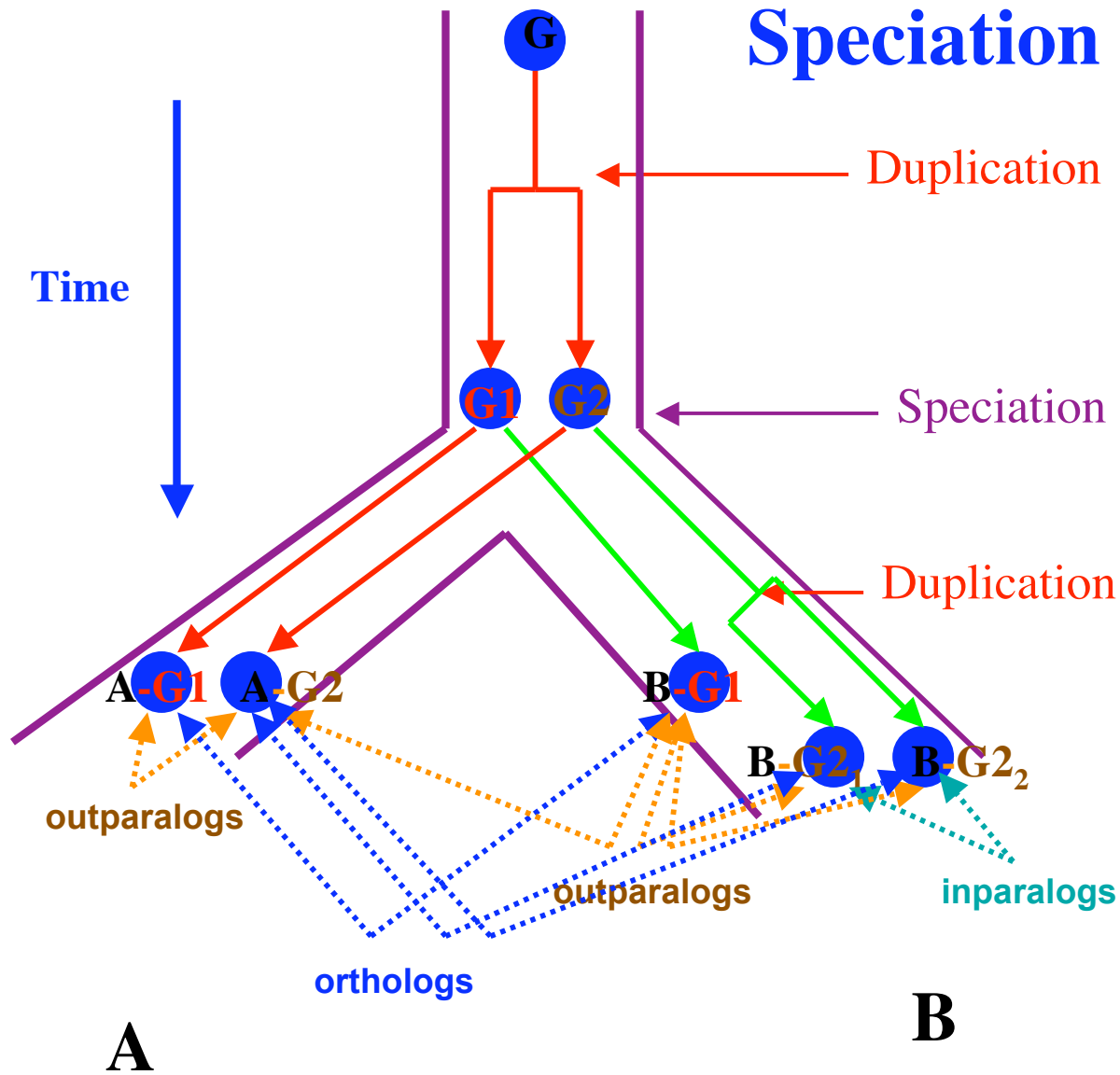
Large scale genome comparisons

- Duplication;**
- Conservation;**
- Specificity (species-specific genes, proteins);**
- Paralogues, orthologues;**
- Families (clusters) of paralogues, of orthologues;**
- Genomes organisations (duplicated, conserved genes);**
- Search for shared motifs in proteins of the same cluster;**
- Protein conservation profiles;**
- Selection pressure analyses**
(synonymous, non synonymous substitutions,..),....

Evolution



Speciation - Duplication

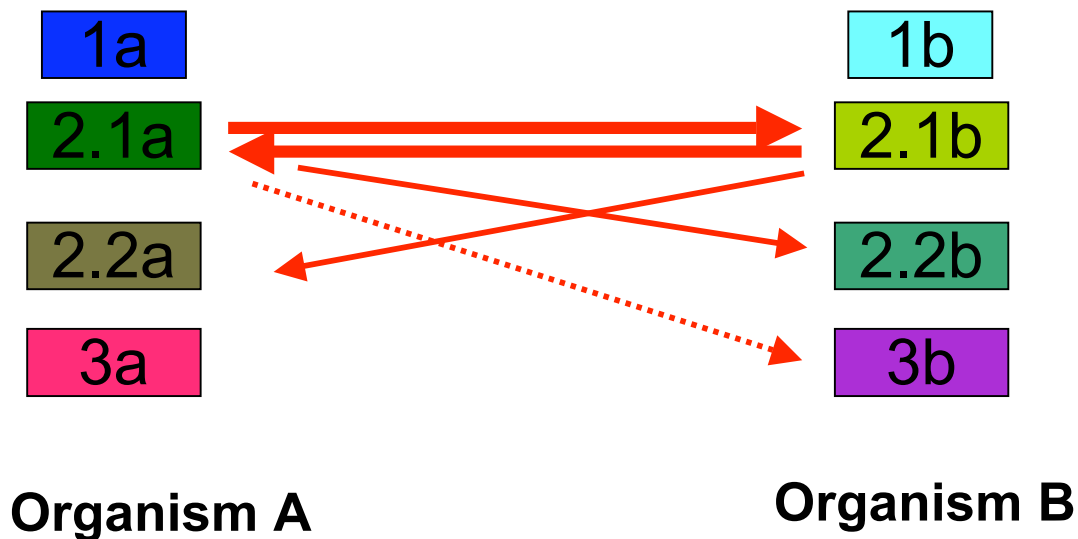


- Speciation
- Duplication
- Inparalogs
- Orthologs
- Outparalogs
- Loss of genes

Predict these events by comparing genomes?

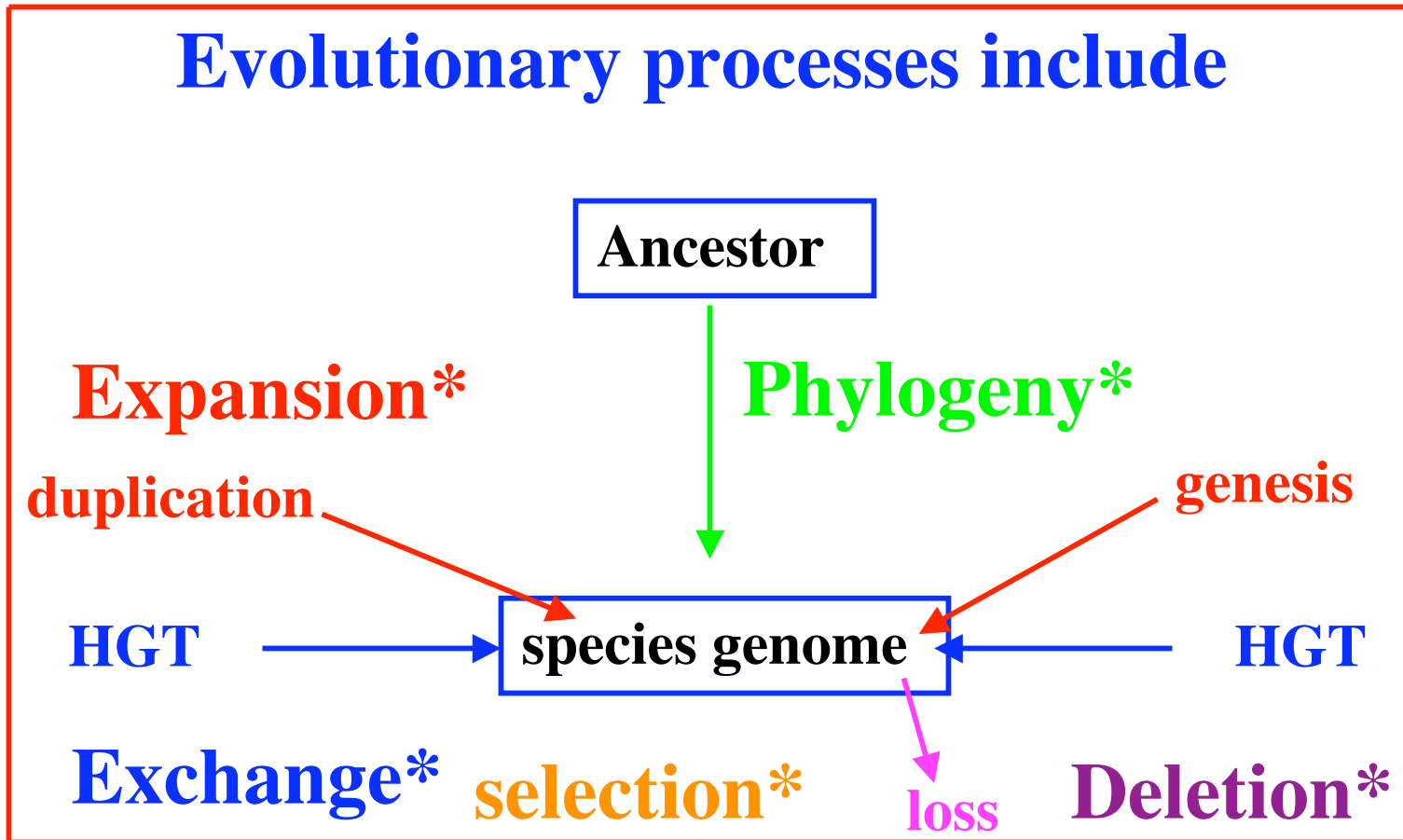
Orthologs / Paralogs

- How to detect orthologous genes?
 - easy way: best reciprocal hit (RBH)



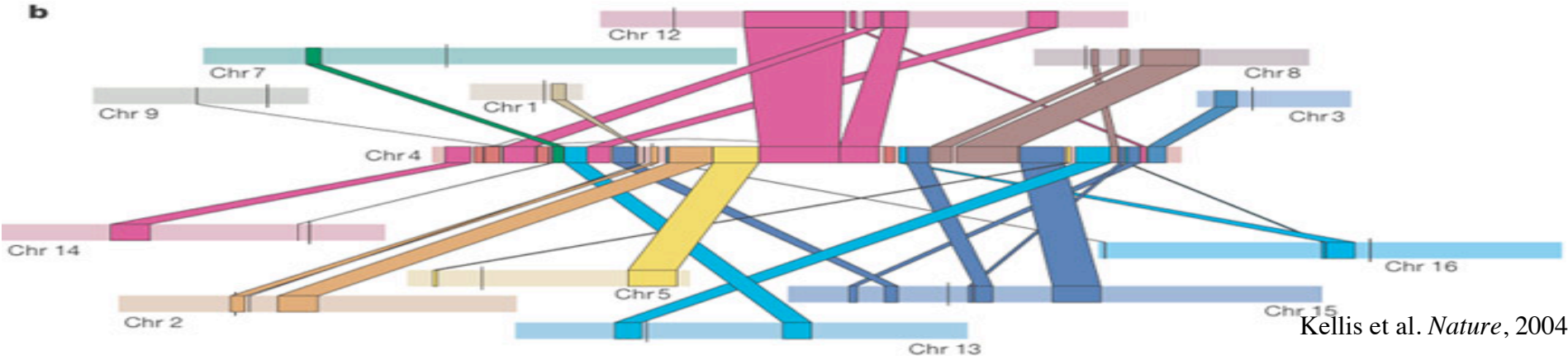
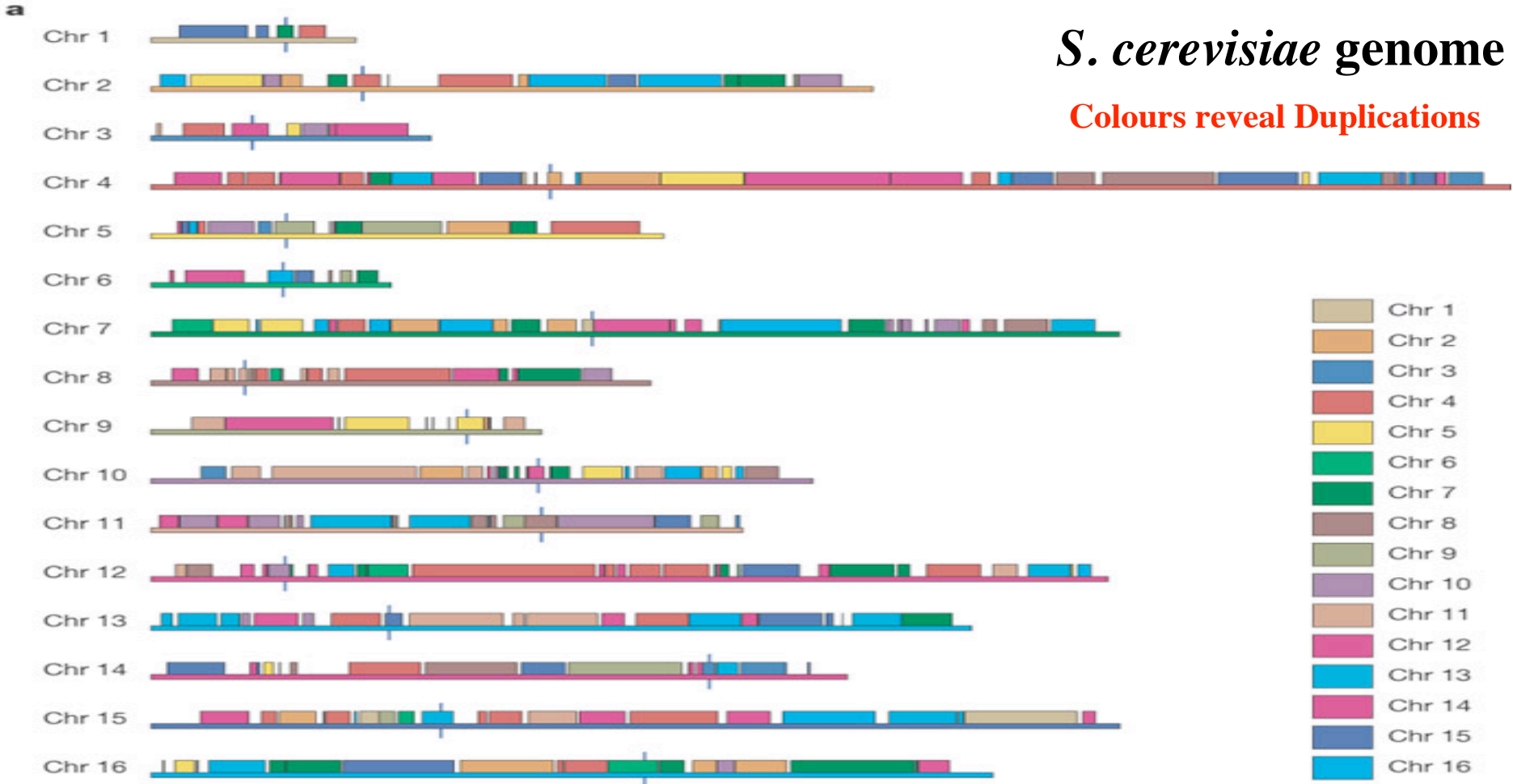
- Large scale comparative analysis of predicted proteomes revealed significant evolutionary processes:

Expansion, **Exchange** and **Deletion**.



S. cerevisiae genome

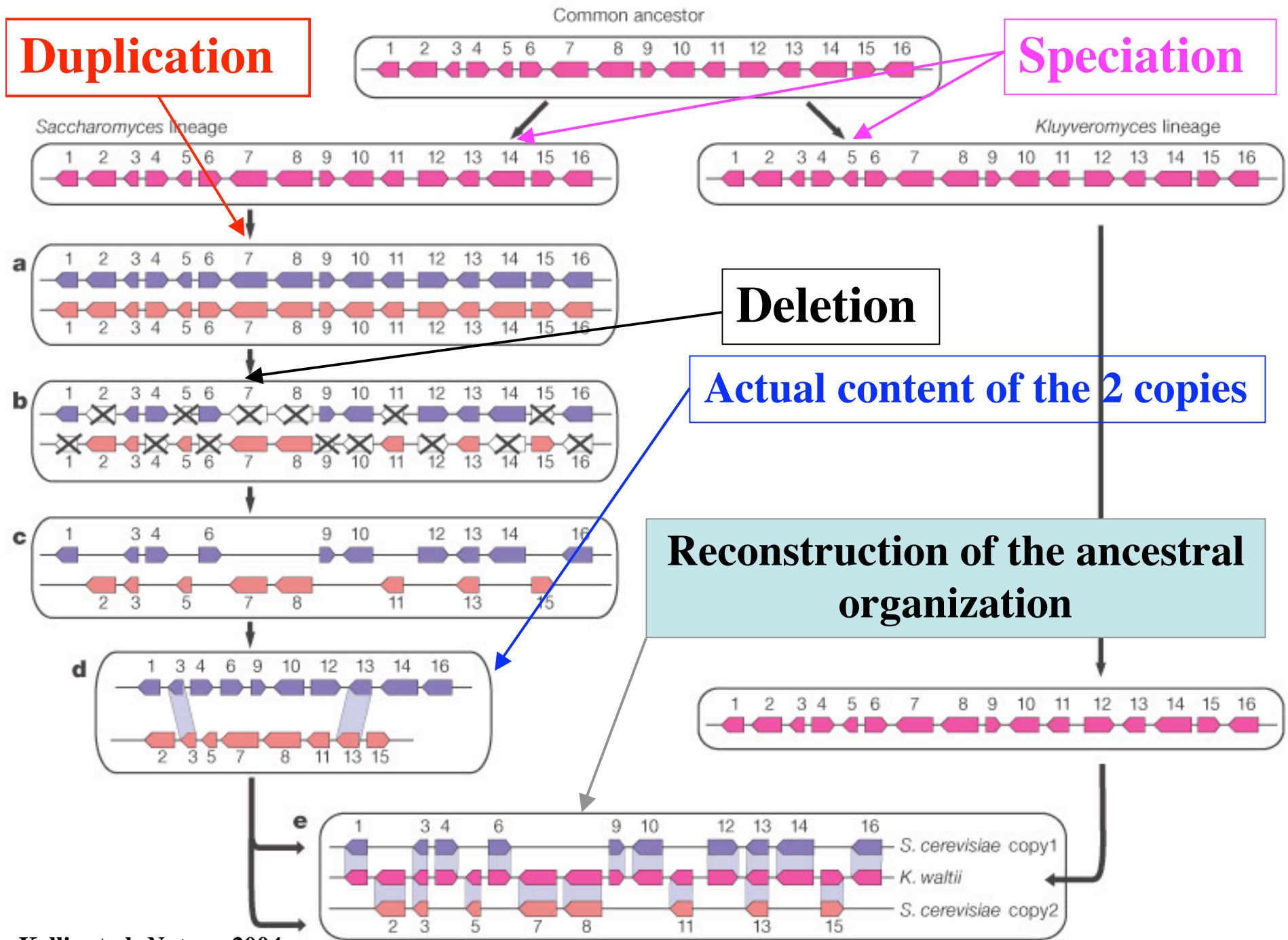
Colours reveal Duplications



Kellis et al. *Nature*, 2004

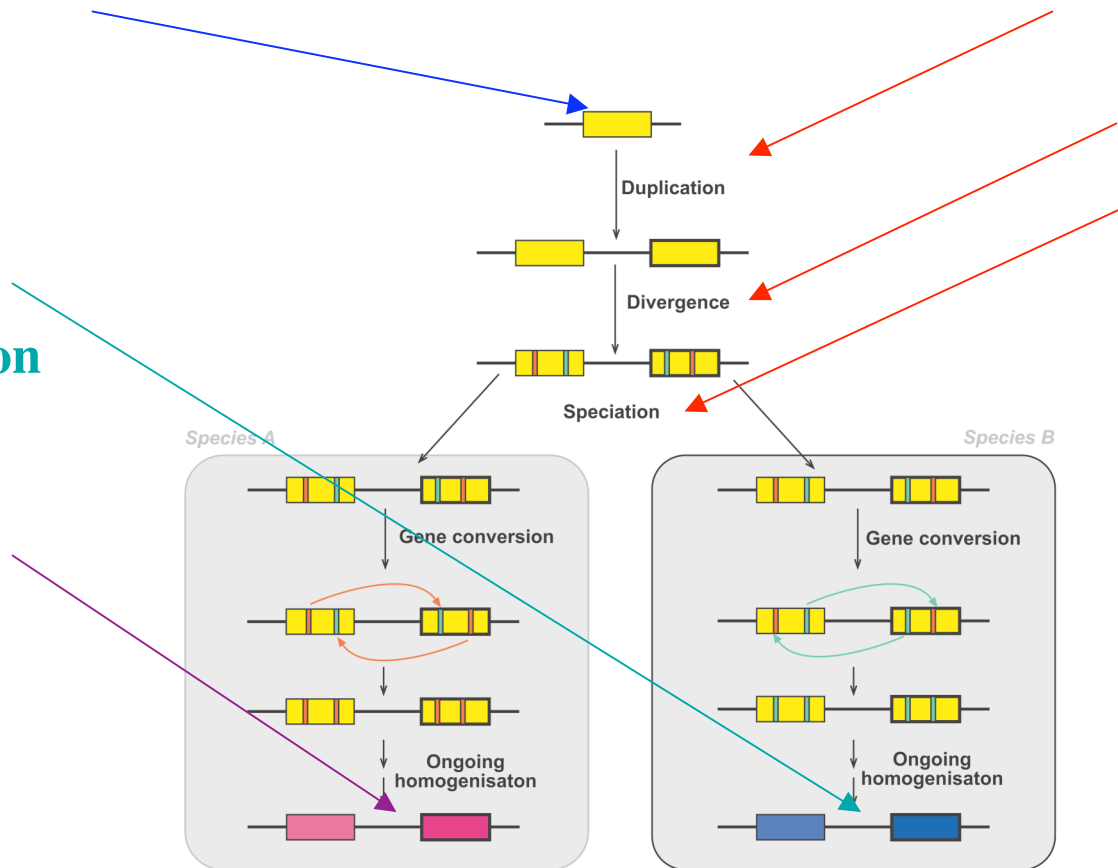
Duplication

Speciation



Original version

Actual version



Search for similarity

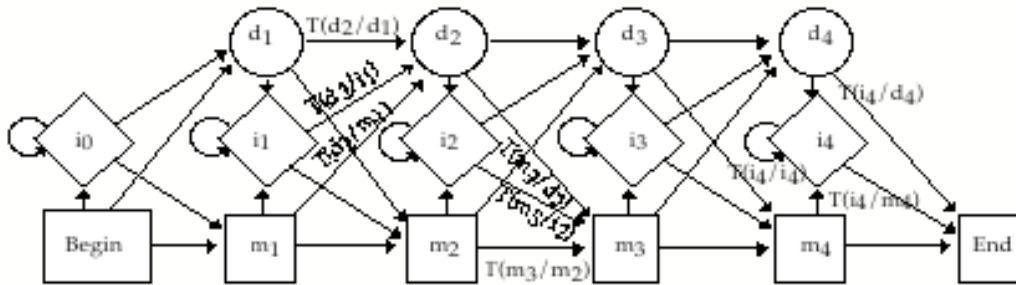
Methods:

- Important to know how algorithms that allow sequence comparisons work,
- There are many comparisons methods,
- Among most used:
- **BLAST**
- **FASTA**
- **Smith-Waterman algorithm**
dynamic programming method
- **HMM (Hidden Markov Model)**

Sequence Comparisons

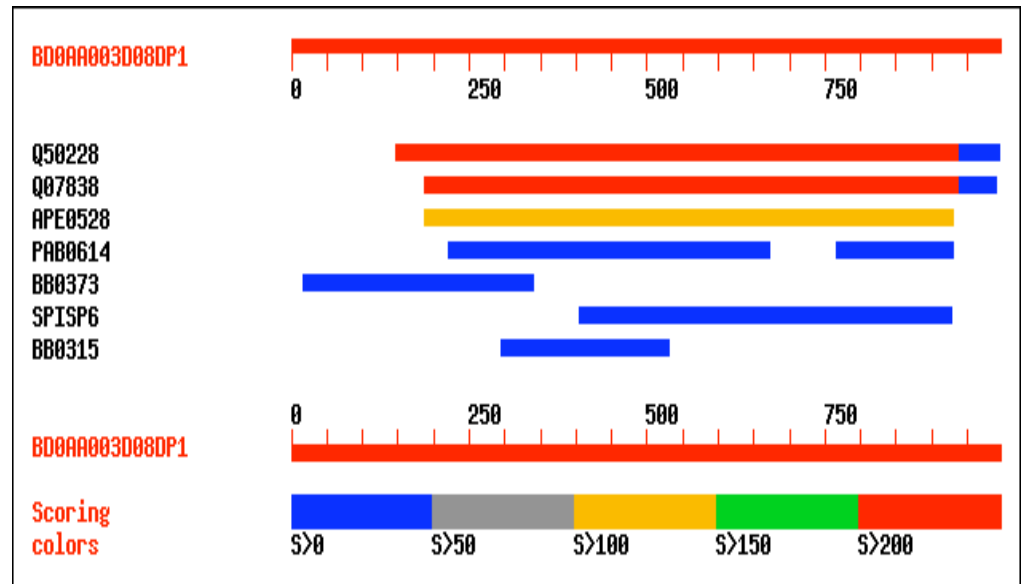
V I T K L G T C V G S
 V I S . . . T Q V G S

V I T K L G T C V G S
 V . S K . G T Q V . S



A linear hidden Markov model

- Identity
- Similarity
- Homology



Comparison of 2 sequences

- **Aims at finding the optimal alignment: the one that shows most similar regions and regions that are less similar.**
- **In describing sequence comparisons, three different terms are commonly used :**

Identity, Similarity and Homology.

➤ **Need for a score that evaluates:**

- **matches**
- **mismatches**
- **gaps**

➤ **and a method that evaluates the numerous possible alignments.**

Identity

- Refers to the occurrence of identical nucleotides or amino acids in the same position in aligned sequences ;
- Identity is objective and well defined;
- Identity can be quantified: Percent i.e the number of identical matches divided by the length of the aligned region.

Similarity

- Sequence similarity takes approximate matches into account, and is meaningful only when such substitutions are scored according to some measure of «difference» with **conservative substitutions assigned more favorable scores than non-conservative ones (substitution matrices)**.
- Given a number of parameters (alphabet, scoring matrix, filtering procedure, etc...), the similarity of an aligned region is defined by a score calculated on that region;
- **The score depends on the chosen parameters;**
- Contrarily to homology : expression like significant or weak similarity are often used.

Homology

- Sequence homology underlies common ancestry and sequence conservation;
- Homology can be inferred, under suitable conditions from sequence similarity ;
- The main objective of sequence similarity searching studies aims at inferring homology between sequences;
- Homology is not a measure.

It is an all or none relationship (i.e homology exists or does not exist. Expressions like : significant or weak homology are meaningless!).

Sequence similarity is a measure of the matching characters in an alignment, whereas homology is a statement of common evolutionary origin.

BLAST

(Basic Local Alignment Search Tool)

Nucleotide BLAST

- Nucleotide query - nucleotide database [[blastn](#)]

Protein BLAST

- Protein query - protein database [[blastp](#)]
- **PSI-BLAST** Position Specific Iterative BLAST

Translated BLAST Searches

- Nucleotide query - Protein db [[blastx](#)]
- Protein query - Translated db [[tblastn](#)]
- Nucleotide query - Translated db [[tblastx](#)]

Search for conserved domains

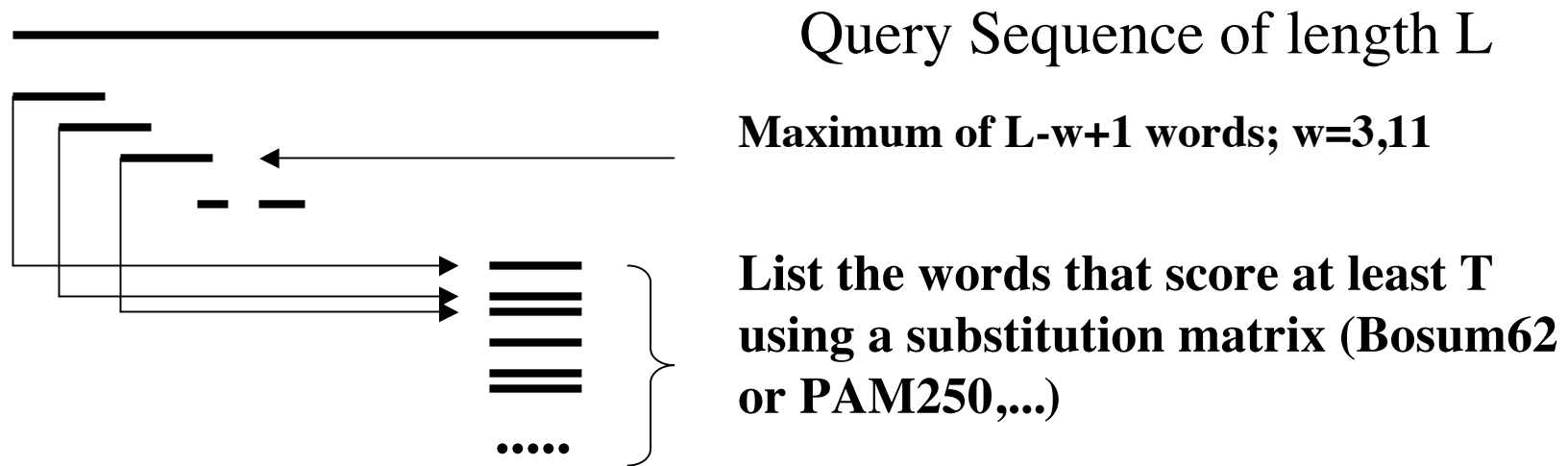
- Search the Conserved Domain Database [[RPS-BLAST](#)]

Pairwise BLAST

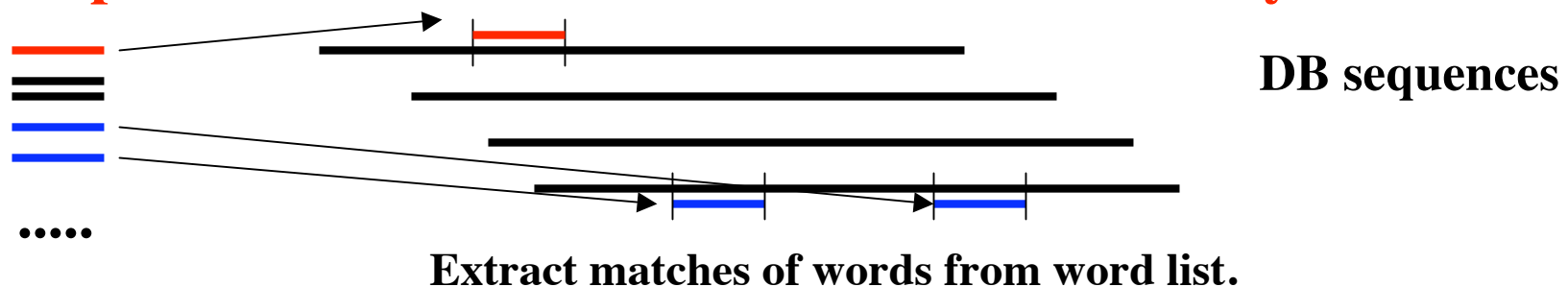
- [BLAST 2 Sequences](#)

Blast algorithm:

(1) Query sequence: list of high scoring words of length w .



(2) Compare the word list to the database and identify exact matches.



(3) For each word match, extend alignment in both directions to find alignments with scores $> S$



BLASTP 2.2.1 [Apr-13-2001]

.....

Query= YAL005c SSA1 heat shock protein of HSP70 family,
cytosolic
(642 letters)

Database: S. cerevisiae proteome version 22/05/2002
5829 sequences; 2,798,770 total letters

.....

Sequences producing significant alignments:		Score (bits)	E Value
YAL005c	SSA1 heat shock protein of HSP70 family, cyt...	674	0.0
YLL024c	SSA2 heat shock protein of HSP70 family, cyt...	663	0.0
YER103w	SSA4 heat shock protein of HSP70 family, cyt...	589	e-169
YBL075c	SSA3 heat shock protein of HSP70 family, cyt...	588	e-169
YJL034w	KAR2 nuclear fusion protein	480	e-136
YDL229w	SSB1 heat shock protein of HSP70 family	428	e-120
YNL209w	SSB2 heat shock protein of HSP70 family, cyt...	427	e-120
YJR045c	SSC1 mitochondrial heat shock protein 70-rel...	336	5e-93
YEL030w	heat shock protein of HSP70 family	324	2e-89
YLR369w	SSQ1 mitochondrial heat shock protein 70	296	4e-81
YBR169c	SSE2 heat shock protein of the HSP70 family	173	7e-44
YPL106c	SSE1 heat shock protein of HSP70 family	172	1e-43
YHR064c	regulator protein involved in pleiotro...	143	6e-35
YKL073w	LHS1 chaperone of the ER lumen	100	4e-22
YLR135w	subunit of SLX1P/Ybr228p-SLX4P complex...	33	0.13

.....

>YLL024c SSA2 P14.1.f13.1 heat shock protein of HSP70 family, cytosolic
Length = 639

Score = 663 bits (2508), Expect = 0.0
Identities = 558/607 (91%), Positives = 570/607 (92%)

```
Query: 1 MSKAVGIDLGTTYSCVAHFANDRVIIANDQGNRTTPSFVAFTDTERLIGDAAKNQAAMN 60
MSKAVGIDLGTTYSCVAHF+NDRVDIIANDQGNRTTPSFV+FTDTERLIGDAAKNQAAMN
Sbjct: 1 MSKAVGIDLGTTYSCVAHFSNDRVDIIANDQGNRTTPSFVGFSTDTERLIGDAAKNQAAMN 60
.....
Query: 601 IMSKLYQ 607
IMSKLYQ
Sbjct: 601 IMSKLYQ 607
```

>YER103w SSA4 P14.1.f13.1 heat shock protein of HSP70 family, cytosolic
Length = 642

Score = 589 bits (2224), Expect = e-169
Identities = 473/609 (77%), Positives = 539/609 (87%), Gaps = 3/609 (0%)

```
Query: 1 MSKAVGIDLGTTYSCVAHFANDRVIIANDQGNRTTPSFVAFTDTERLIGDAAKNQAAMN 60
MSKAVGIDLGTTYSCVAHFANDRV+IIANDQGNRTTPS+VAFTDTERLIGDAAKNQAAMN
Sbjct: 1 MSKAVGIDLGTTYSCVAHFANDRVEIIANDQGNRTTPSYVAFTDTERLIGDAAKNQAAMN 60
.....
Query: 598 ANPIMSKLY 606
ANPIMSK+Y
Sbjct: 601 ANPIMSKFY 609
```

>YBL075c SSA3 P14.1.f13.1 heat shock protein of HSP70 family, cytosolic
Length = 649

Score = 588 bits (2220), Expect = e-169
Identities = 467/609 (76%), Positives = 539/609 (87%), Gaps = 3/609 (0%)

```
Query: 1 MSKAVGIDLGTTYSCVAHFANDRVIIANDQGNRTTPSFVAFTDTERLIGDAAKNQAAMN 60
MS+AVGIDLGTTYSCVAHF+NDRV+IIANDQGNRTTPS+VAFTDTERLIGDAAKNQAA+N
Sbjct: 1 MSRAVGIDLGTTYSCVAHFSNDRVEIIANDQGNRTTPSYVAFTDTERLIGDAAKNQAAMN 60
.....
Query: 598 ANPIMSKLY 606
ANPIM+K+Y
Sbjct: 601 ANPIMTKFY 609
```

>YJL034w KAR2 P14.1.f13.1 nuclear fusion protein
Length = 682

.....