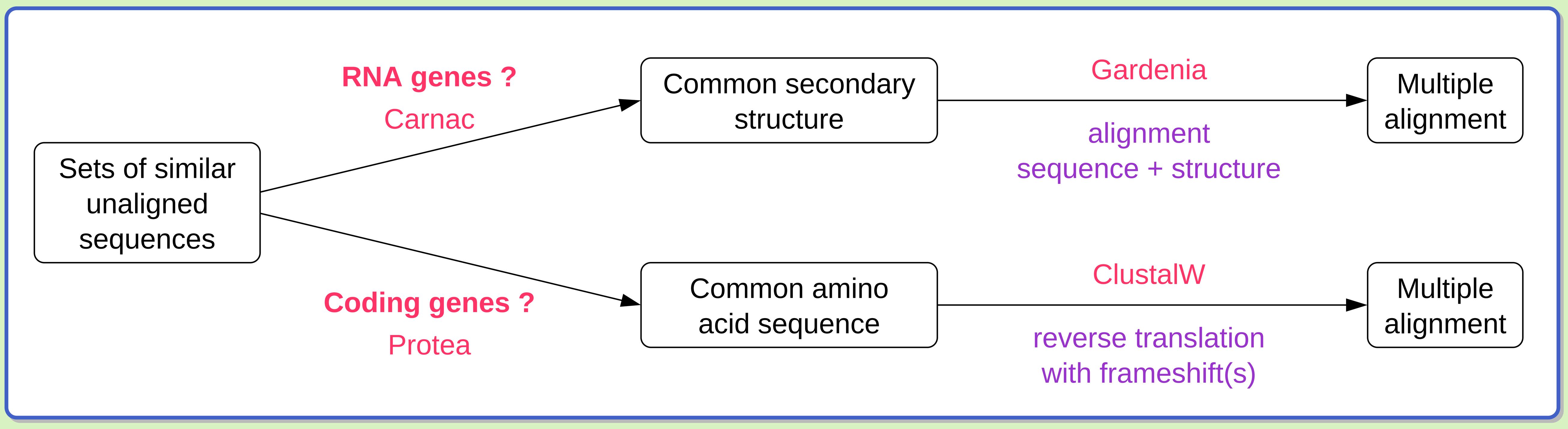
Identification of protein-coding genes and RNA genes by comparative analysis

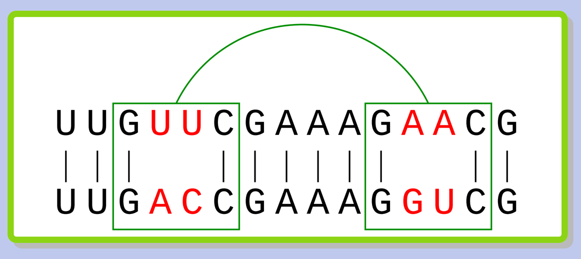
Arnaud Fontaine and Hélène Touzet
LIFL (UMR USTL/CNRS 8022) – Sequoia (INRIA Futurs)
{arnaud.fontaine, helene.touzet}@lifl.fr

From function to alignment



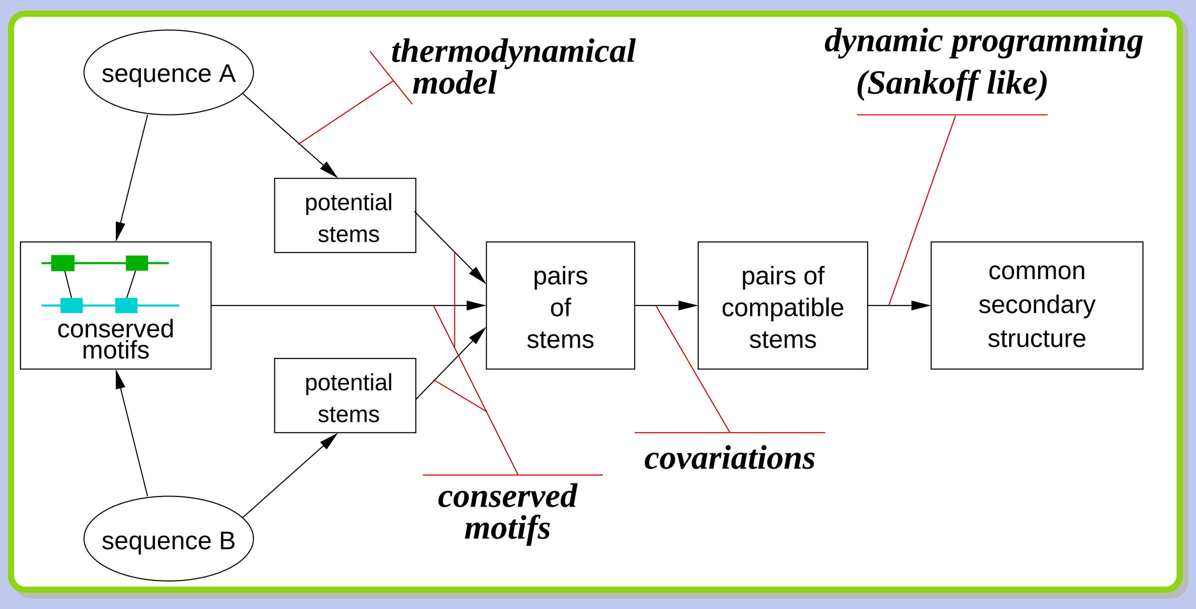
Carnac: identification of RNA genes

Evolutionary pattern



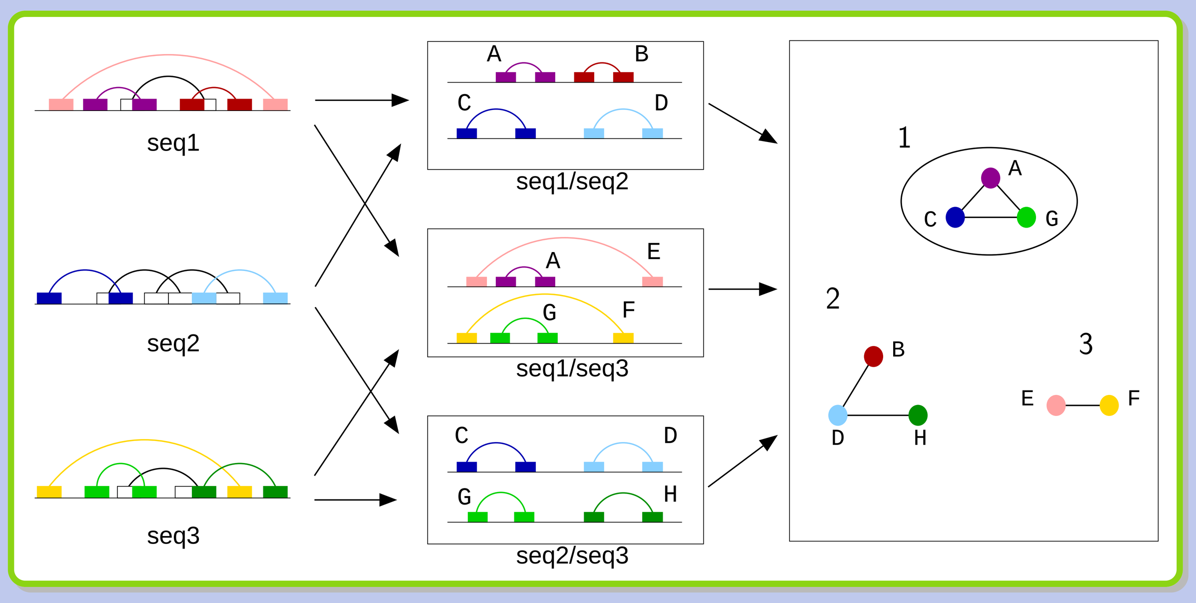
2 sequences

Looking for a conserved secondary structure



From 2 to n sequences

Stem graph



Example

- 5 RNA sequences from BRAliBase [1] database (tRNA sequences)
- Prediction of common secondary structure with Carnac
- Multiple alignment with Gardenia

Reference alignment from BRAliBase

```
GGAGAGATGGCTGAGT-GGACTAAAGCGGCGGATTGCTAATCCGTT-GTACAATTTTTTTGTAC--CGAGGGTTCGAATCCCTCTCTTTCCG
GGGGGTGTAGCTCAGT-GGT--AGAGCGCATGCTTTGCATGTATGAG-----------------GTCTTGGGTTCAATCCCCAGCATCTCCA
GCGCCCGTAGCTCAGCTGGAT-AGAGCGCTGCCCTCCGGAGGCAGAG-----------------GTCACAGGTTCGAATCCTGTCGGGCGCA
GGGTTATTAGCTCAGTTGGTT-AGAGCACACCCCTGATAAGGGTGAG-----------------GTCCCTGGTTCAAATCCAGGATAACCCA
TCCTCAGTAGCTCAGT-GGT--AGAGCGGTCGGCTGTTAACCGATTG-----------------GTCGTAGGTTCGAATCCTACTTGGGGAG
```

ClustalW alignment

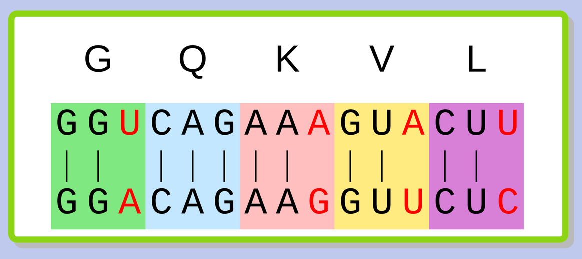
```
GGAGAGATGGCTGAGTGGACTAAAGCGGCGGATTGCTAATCCGTTGTACAATTTTTTTGTACCGAGGGTTCGAATCCCTCTCTTTCCG
GGGGGTGTAGCTCAGT--GGTAGAGCGCATGCTTTGCATGTATGAGGTCTTGGGTTCAATCCCCAGCATCTCCA--------------
GCGCCCGTAGCTCAGCTGGATAGAGCGCTGCCCTCCGGAGGCAGAGGTCACAGGTTCGAATCCTGTCGGGCGCA--------------
GGGTTATTAGCTCAGTTGGTTAGAGCACACCCCTGATAAGGGTGAGGTCCCTGGTTCAAATCCAGGATAACCCA--------------
TCCTCAGTAGCTCAGT--GGTAGAGCGGTCGGCTGTTAACCGATTGGTCGTAGGTTCGAATCCTACTTGGGGAG--------------
```

Gardenia alignment on Carnac prediction

```
--GGAGAGATGGCTGAG-TGGACTAAAGCGGCGGATTGCTAATCCGTTGTACAATTTTTTTGTACCGAGGGTTCGAATCCCTC---TCTTTCCG
GGGGG-TG-TAGCTCAG-TGG--TAGAGCGCATGCTTTGCATGTATGAGG----T---------C-TTGGGTTC-AATCCCCAGCATCT--CCA
GCGCC-CG-TAGCTCAGCTGGA-TAGAGCGCTGCCCTCCGGAGGCAGAGG----T---------C-ACAGGTTCGAATCCTGT-CGGGC--GCA
GGGTT-AT-TAGCTCAGTTGGT-TAGAGCACACCCCTGATAAGGGTGAGG----T---------C-CCTGGTTCAAATCCAG-GATAAC--CCA
TCCTC-AG-TAGCTCAG-TGG--TAGAGCGGTCGGCTGTTAACCGATTGG----T---------C-GTAGGTTCGAATCCTAC-TTGGG--GAG
(((((                       ((((((      ))))))                    (((((      )))))   )))  ))
```

Protea: identification of protein-coding genes

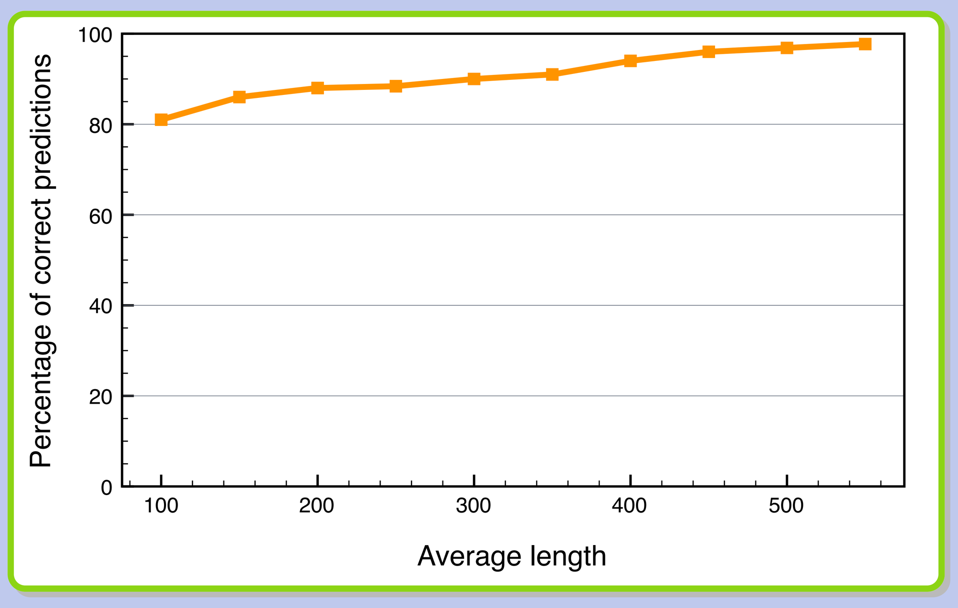
Evolutionary pattern



2 sequences

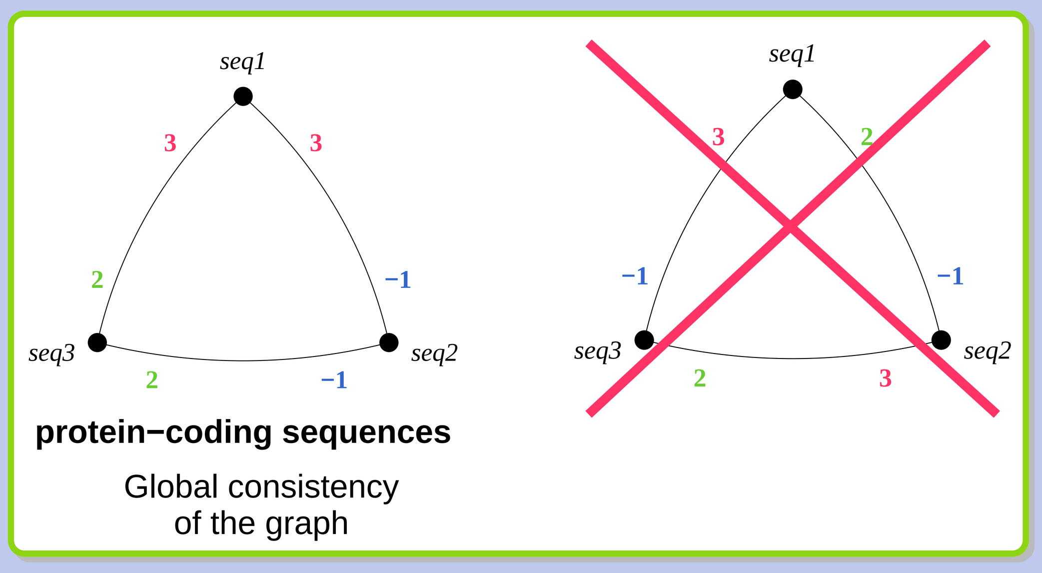
Looking for conserved reading frames

Pandit [2] database: among the 36 possible reading frame pairs, the real one is the most conserved



From 2 to n sequences

Reading frame consistency graph



Example

- 3 homologous coding sequences from Pandit [2] database (PF07974 family - EGF-like domain)
- Prediction of correct reading frames with Protea
- Reverse translation of multiple alignment build by ClustalW [3] on predicted amino acid sequences

Reference alignment from Pandit

```
TGCAGCCCCCGGGAGGGCCAGCCCGCCTGCAGCCAGCGGGGCGAGTGCCTG------TGTGGCCAATGTGTCTGCCATAGCAGTGACTTTGGCAAGATCACGGGCAAGTACTGC
TGCCGGTCACCTGAAAACAACGAAATCTGCAGTGGAAACGGACAATGTGTA------TGTGGACAATGTATGTGTAACTCTGACGATGACCGCCACTATAGTGGCAAATACTGC
TGTTTTGGAAAAGGATCC---------TGTCATGGAGATGGAAGCCGCGAAGGCAGT---GGAAAGTGTAAATGTGAGACTGGA------------TATACTGGAAATCTATGC
```

ClustalW alignment

```
TGCAGCCCCCGGGAGGGCCAGCCCGCCTGCAG-CCAGCGGGGCGAGTGCCTGTGTGGCCAATGTGTCTGCCATAGCAGTGACTTTGGCAAGATCACGGGCAAGTACTGC
TGCCGGTCACCTGAAAACAACGAAATCTGCAG-TGGAAACGGACAATGTGTATGTGGACAATGTATGTGTAACTCTGACGATGACCGCCACTATAGTGGCAAATACTGC
----TGTTTTGGAAAAGGATCCTGTCATGGAGATGGAAGCCGCGAAGGCA---GTGGAAAGTGTAAATGTGAGACTGGATATACTGGAAATCTATGC------------
```

Protea alignment

```
TGCAGCCCCCGGGAGGGCCAGCCCGCCTGCAGCCAGCGGGGC---GAGTGCCTGTGTGGCCAATGTGTCTGCCATAGCAGTGACTTTGGCAAGATCACGGGCAAGTACTGC
TGCCGGTCACCTGAAAACAACGAAATCTGCAGTGGAAACGGA---CAATGTGTATGTGGACAATGTATGTGTAACTCTGACGATGACCGCCACTATAGTGGCAAATACTGC
TGTTTT---------GGAAAAGGATCCTGTCATGGAGATGGAAGCCGCGAAGGCAGTGGAAAGTGTAAATGTGAGACTGGA------------TATACTGGAAATCTATGC
```

For experimental results, see our papers
Protea [4], Carnac [5, 6] and Gardenia [7] are freely available from our web server

http://bioinfo.lifl.fr

References

[1] Paul P. Gardner, Andreas Wilm, and Stefan Washietl. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucl. Acids Res.*, 33(8):2433–2439, 2005.

[2] Simon Whelan, Paul I. W. de Bakker, Emmanuel Quevillon, Nicolas Rodriguez, and Nick Goldman. PANDIT: an evolution-centric database of protein and associated nucleotide domains with inferred trees. *Nuc. Ac. Res.*, 34:D327–D331, 2006.

[3] JD Thompson, DG Higgins, and TJ Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nuc. Ac. Res.*, 22(22):4673–4680, 1994.

[4] A. Fontaine and H. Touzet. Computational identification of protein-coding sequences by comparative analysis. In *Proceedings of the 1st IEEE international conference on Bioinformatics and Biomedecine (BIBM), Silicon Valley, California*, 2007.

[5] H. Touzet and O. Perriquet. CARNAC: folding families of related RNAs. *Nucleic Acids Research*, 32 (Supplement 2):142–145, 2004.

[6] H. Touzet. *Comparative analysis of RNA genes: the CaRNAc software*, volume Methods in Molecular Biology, Special issue on comparative genomics I, pages 465–473. Humana Press, 2007.

[7] G. Blin and H. Touzet. How to compare arc-annotated sequences: The alignment hierarchy. In *13th International Symposium on String Processing and Information Retrieval (SPIRE)*, volume 4209 of *Lecture Notes in Computer Science*, pages 291–303. Springer Verlag, 2006.