Comparing Sequences and Trees From Computational Biology to Music Analysis

Julien Allali, Pascal Ferraro, Pierre Hanna and Matthias Robine

PIMS - CNRS, University of Bordeaux 1, LaBRI, SIMBALS

Comparison of Biological Structures

- What do we compare ?
  - DNA (coding and non coding regions)
  - RNA
  - proteins
  - Plant Architecture
- Data
  - Sequences or trees of nucleotides
  - Sequences of amino-acids
  - Sequences or trees of elementary entities

  $\Longrightarrow$ **Comparison of strings (or trees) of characters**
- Why do we compare ?
  - Search for similar biological functions,
  - Identification of comparable structures,
  - Construction of phylogenetic trees,
  - Identification of gene mutation,
  - Detection of gene transfer.

Comparison of Biological Structures

- What do we compare ?
  - DNA (coding and non coding regions)
  - RNA
  - proteins
  - Plant Architecture
- Data
  - Sequences or trees of nucleotides
  - Sequences of amino-acids
  - Sequences or trees of elementary entities

$\Longrightarrow$ **Comparison of strings (or trees) of characters**

- Why do we compare ?
  - Search for similar biological functions,
  - Identification of comparable structures,
  - Construction of phylogenetic trees,
  - Identification of gene mutation,
  - Detection of gene transfer.

Comparison of Biological Structures

- What do we compare ?
  - DNA (coding and non coding regions)
  - RNA
  - proteins
  - Plant Architecture
- Data
  - Sequences or trees of nucleotides
  - Sequences of amino-acids
  - Sequences or trees of elementary entities

  $\Longrightarrow$ **Comparison of strings (or trees) of characters**

- Why do we compare ?
  - Search for similar biological functions,
  - Identification of comparable structures,
  - Construction of phylogenetic trees,
  - Identification of gene mutation,
  - Detection of gene transfer.

Measure of Musical Similarity

- What do we compare ?
  - Timbre
  - Rhythm
  - Melodies
- Database
  - Audio (wav, mp3, . . .)
  - Symbolic (MIDI)

  $\Longrightarrow$ **Symbolic melodic similarity** = Comparison of sequences (or trees) of notes
- Why do we compare?
  - Music Information Retrieval,
  - Search for similarities in musical database,
  - Automatic detection of plagiarism,
  - Musical analysis by self-similarity.

Measure of Musical Similarity

- What do we compare ?
  - Timbre
  - Rhythm
  - Melodies
- Database
  - Audio (wav, mp3, . . .)
  - Symbolic (MIDI)

$\Longrightarrow$ **Symbolic melodic similarity** = Comparison of sequences (or trees) of notes

- Why do we compare?
  - Music Information Retrieval,
  - Search for similarities in musical database,
  - Automatic detection of plagiarism,
  - Musical analysis by self-similarity.

Measure of Musical Similarity

- What do we compare ?
  - Timbre
  - Rhythm
  - Melodies
- Database
  - Audio (wav, mp3, . . .)
  - Symbolic (MIDI)

  $\Longrightarrow$ **Symbolic melodic similarity** = Comparison of sequences (or trees) of notes
- Why do we compare?
  - Music Information Retrieval,
  - Search for similarities in musical database,
  - Automatic detection of plagiarism,
  - Musical analysis by self-similarity.

Outline

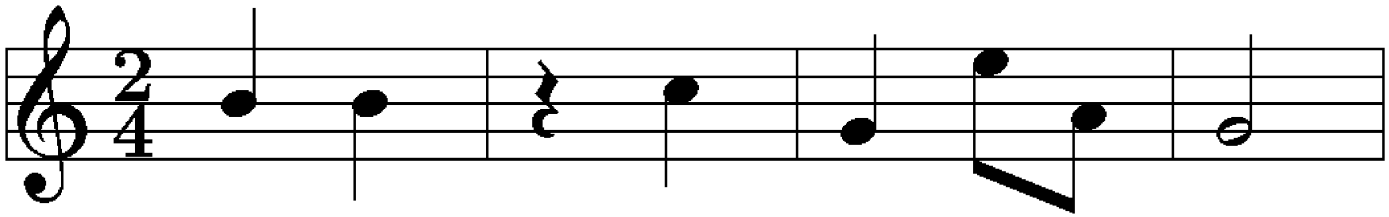
Outline

Molecular Sequences

- Molecule of DNA or RNA : linear suite of nucleotides = primary structure
- DNA : a molecule is always made of a sugar, a phosphate group and one of the four nucleic acids: Adenine, Cytosine, Guanine and Thymine. There are represented by an alphabet made of their initials : $\{A, C, G, T\}$
- RNA : Thymine T is replaced by Uracil U.
- Sometimes, some positions in the sequence are unknown $\Rightarrow$ an extended alphabet is used.
- Proteins : sequences of amino-acids (20 characters in the alphabet).

Music Representation

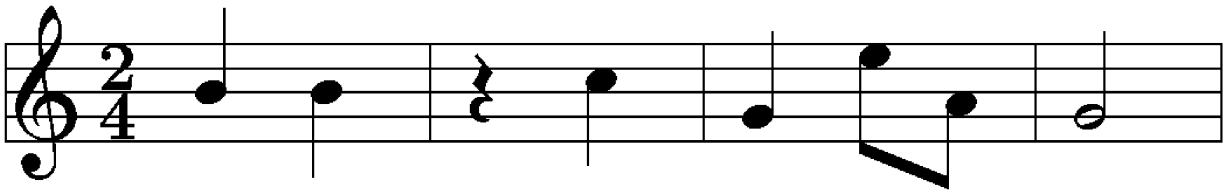
Melody: **sequence** of notes represented by their pitch and duration [Mongeau and Sankoff, 1990].



represented by the sequence :

```
(B4 B4 r4 C4 G4 E2 A2 G8)
```

Different Alphabets



- Absolute pitch : Exact pitch in MIDI notation
  `59 59 - 60 55 64 57 55`
- Contour : Up, Down, Same
  `- S - U D U D D`
- Interval : number of half-tones with previous note (Modulo 12, oriented)
  `- 0 - +1 -5 +3 -5 -2`
- Difference between the current note and the key (Modulo 12, oriented)
  $\Rightarrow$ Problem to determine the correct key
  `-1 -1 - . -7 +4 -9 -7`
- Representation of duration:
  - Absolute duration,
  - Contour,
  - Interval

Similarity Between Sequences of Symbols

$\Rightarrow$ *String matching* Algorithms

Edit Operations :

- Insertion (I)
- Deletion (D)
- Matching (M)
- Substitution (S)

Example:
distance(APPLIED,PRINCE) ?

word 1	A	P	P	L	I	_	_	E	D
word 2	_	P	R	_	I	N	C	E	_
operation	D	M	S	D	M	I	I	M	D

Edit-Distance : Local Alignment

[Smith and Waterman 1981]

- Determine the best alignment between 2 sequences
- A cost is assigned to each edit operation. For example:
  - Deletion/Insertion : -2
  - Substitution : -1
  - Matching : 1
- Dynamic Programming Algorithm
- Output :
  - What is the best score ?
  - What are the positions corresponding to the best alignment ?
- In local alignment only alignment with a positive score are kept.

Local Alignment

	_	P	R	I	N	C	E	S	S
_	0	0	0	0	0	0	0	0	0
R	0								
I	0								
C	0								
E	0								

$$M[i,j] = \max \begin{cases} 0 \\ M[i-1,j] - 2 \\ M[i,j-1] - 2 \\ M[i-1,j-1] + \text{match}(\text{word 1}[i], \text{word 2}[j]) \end{cases}$$

Local Alignment

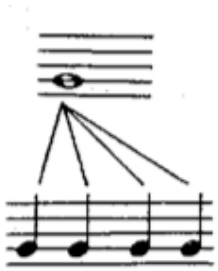
	_	P	R	I	N	C	E	S	S
_	0	0	0	0	0	0	0	0	0
R	0	0	1	0	0	0	0	0	0
I	0	0	0	**2**	0	0	0	0	0
C	0	0	0	0	1	1	0	0	0
E	0	0	0	0	0	0	**2**	0	0

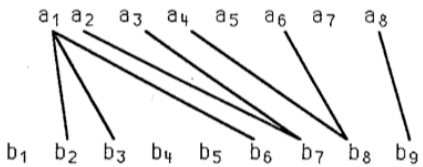
$\Longrightarrow$ Similarity score = 2

corresponding to the alignment :

P	R	I	N	C	E	S	S
_	R	I	_	C	E	_	_

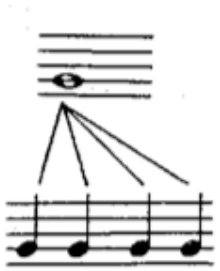
Adaptation to Music

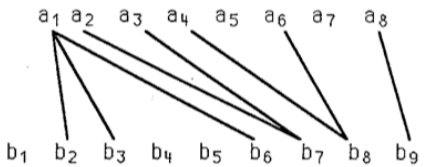




- Introduction of new edit operations: merge and split (Mongeau and Sankoff, 1990)
- Definition of accurate edit operation costs (Ferraro and Hanna, 2007), (Robine et al., 2008)
- Local transpositions (Allali et al., 2008)

Adaptation to Music

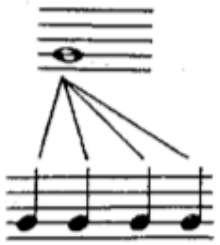


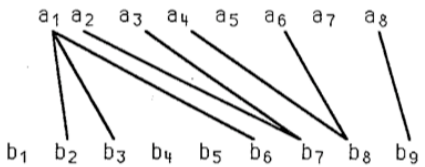


- Introduction of new edit operations: merge and split (Mongeau and Sankoff, 1990)

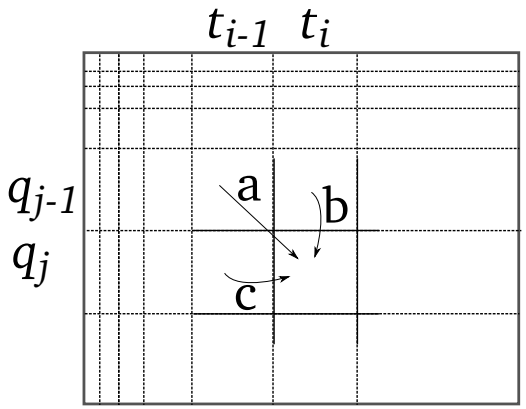
- Definition of accurate edit operation costs (Ferraro and Hanna, 2007), (Robine et al., 2008)
- Local transpositions (Allali et al., 2008)

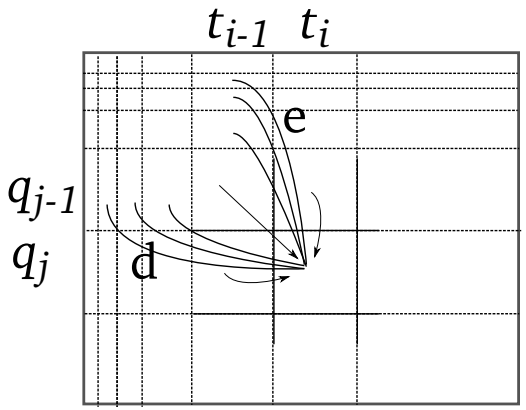
Adaptation to Music





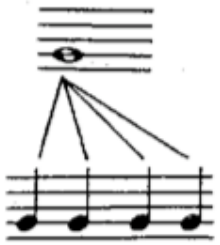
- Introduction of new edit operations: merge and split (Mongeau and Sankoff, 1990)

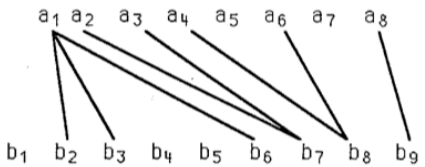




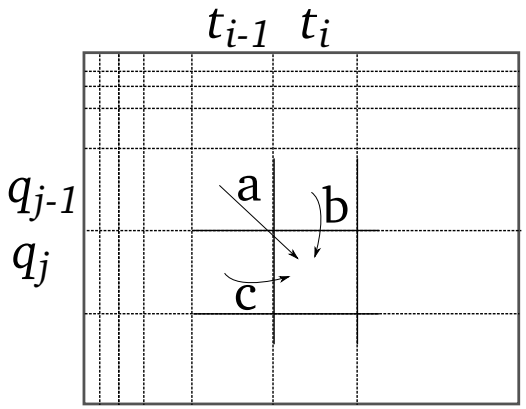
- Definition of accurate edit operation costs (Ferraro and Hanna, 2007), (Robine et al., 2008)
- Local transpositions (Allali et al., 2008)

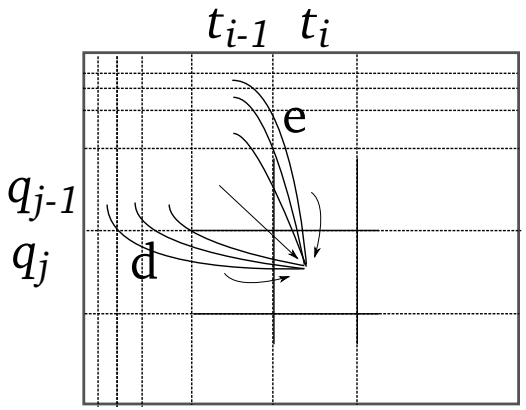
Adaptation to Music





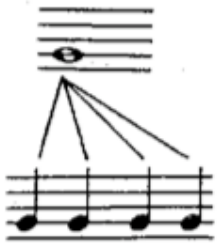
- Introduction of new edit operations: merge and split (Mongeau and Sankoff, 1990)

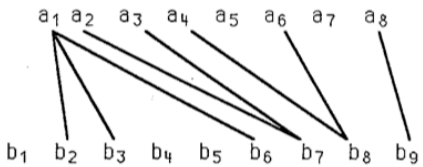




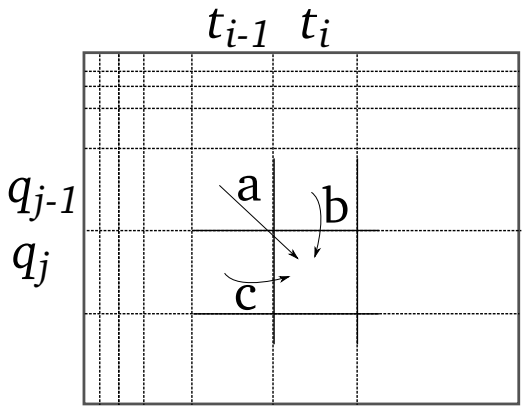
- Definition of accurate edit operation costs (Ferraro and Hanna, 2007), (Robine et al., 2008)
- Local transpositions (Allali et al., 2008)

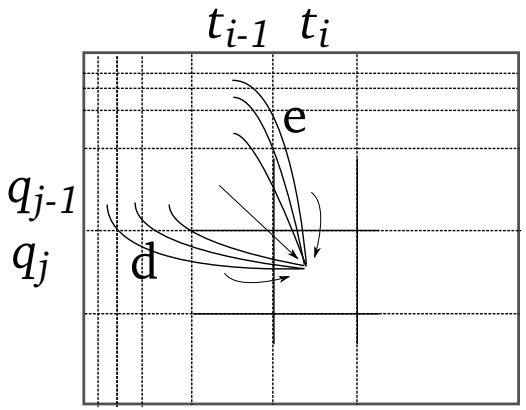
Adaptation to Music





- Introduction of new edit operations: merge and split (Mongeau and Sankoff, 1990)



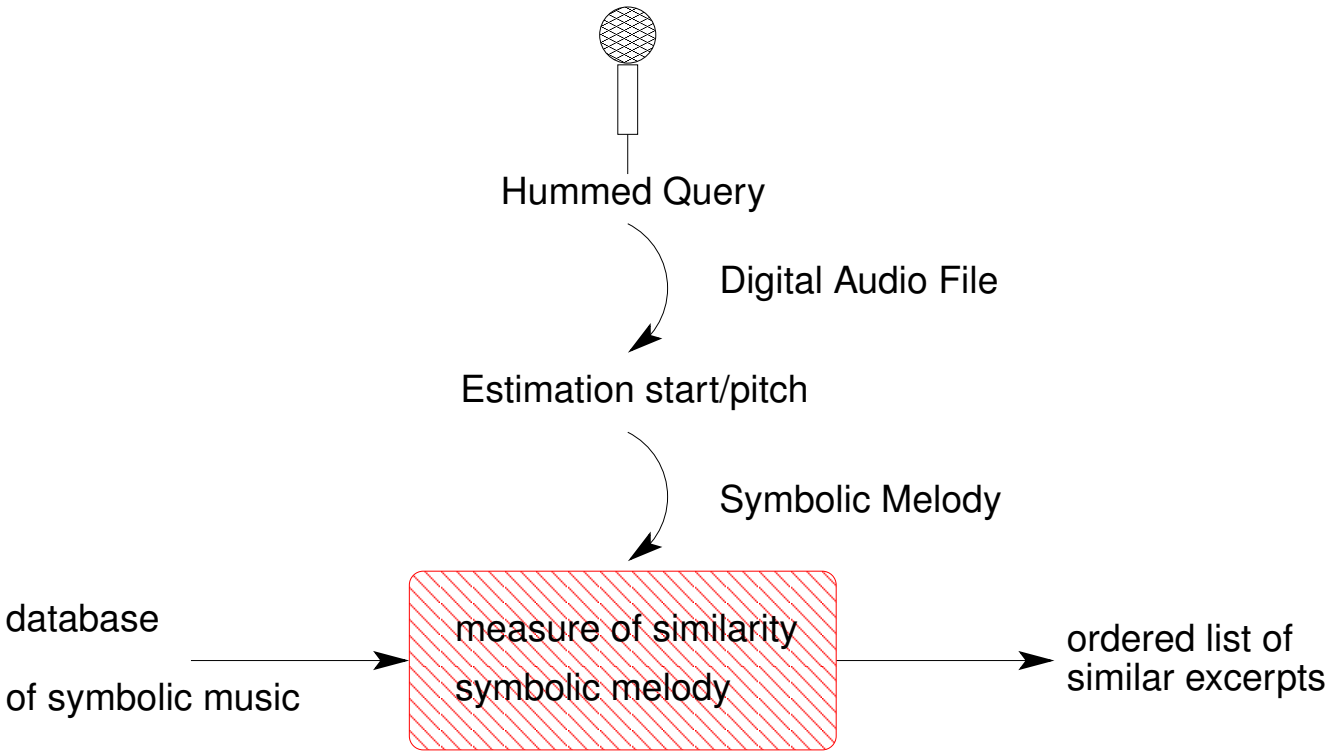


- Definition of accurate edit operation costs (Ferraro and Hanna, 2007), (Robine et al., 2008)
- Local transpositions (Allali et al., 2008)

Applications - *Music Information Retrieval*

- Research by similarity in musical database,
- Query-by-humming,
- Automatic detection of plagiarism,
- Musical analysis by self-similarity.
- ...

Query by Humming



- *Example 1* : Monophonic Query $\Rightarrow$ Smoke on the Water
- *Example 2* : Monophonic Query $\Rightarrow$ First extracted excerpt

Plagiarism Detection

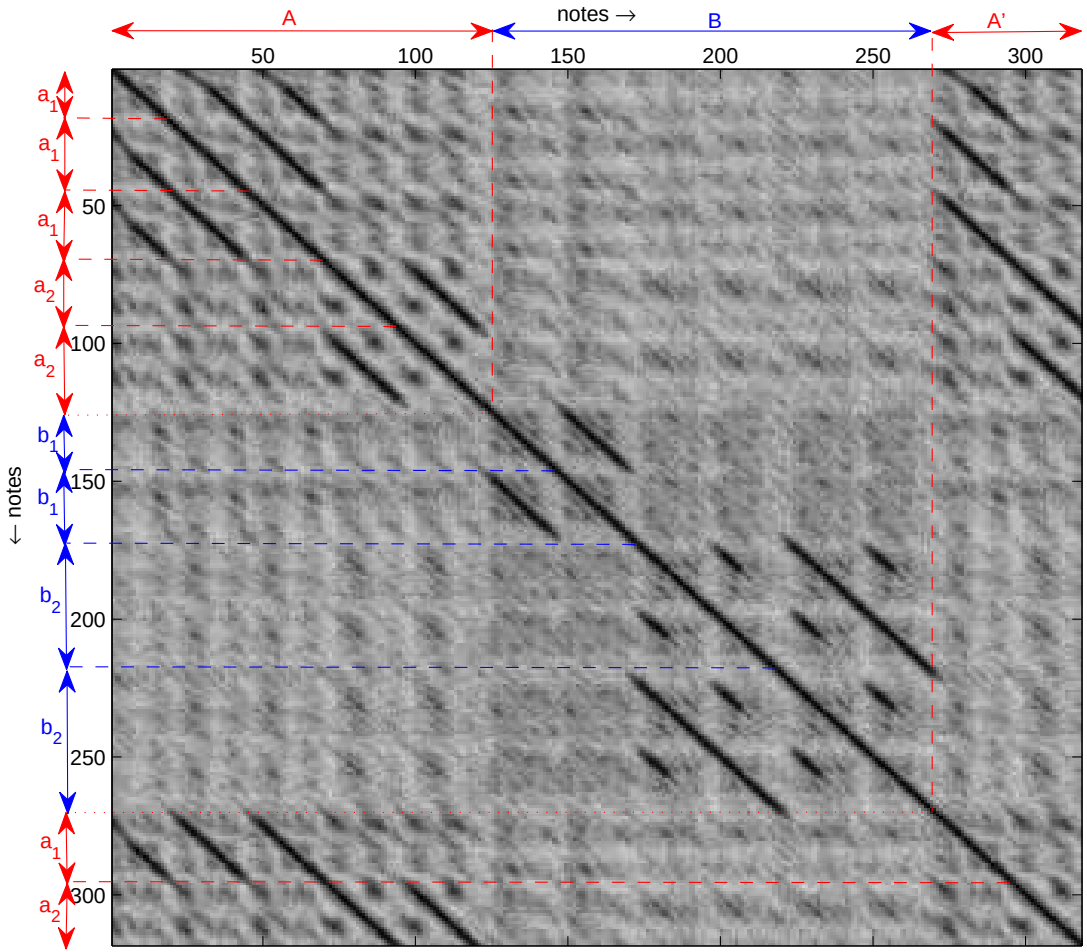
Query	**rank** 1 **score** 1	**rank** 2 **score** 2	**rank** 3 **score** 3
Heim vs Universal (1946)			
Vagyok	Vagyok 248.6	**Perhaps** 123.5	X 92.8
Perhaps	Perhaps 215.5	**Vagyok** 123.5	X 76.8
R. Mack vs G. Harrison (1976)			
Sweet Lord	Sweet Lord 178.9	**So Fine** 83.0	X 77.5
So Fine	So Fine 199.7	**Sweet Lord** 83.0	X 75.3
Selle vs Gibb (1984)			
Let It End	Let It End 192.4	**How Deep** 118.1	X 68.9
How Deep	How Deep 202.8	**Let It End** 118.1	X 83.8

Results on a database of 1650 excerpts (Robine et al., 2007)

Musical Analysis by Self-Similarity (Hanna, Robine, Ferraro, 2008)

- Goal: visualization of repetitions inside an excerpt
- Method:
  - Decomposition of the excerpt into a suite of fixed size segments
  - Similarity Measure between two consecutive segments
  - Normalized score (grey levels)
- Example : Visualisation of the ABA musical structure of the *menuet* of the *Water Music Suite No.1 en F* by Haendel

Musical Analysis by Self-Similarity



Water Music Suite No.1 en F, Haendel.

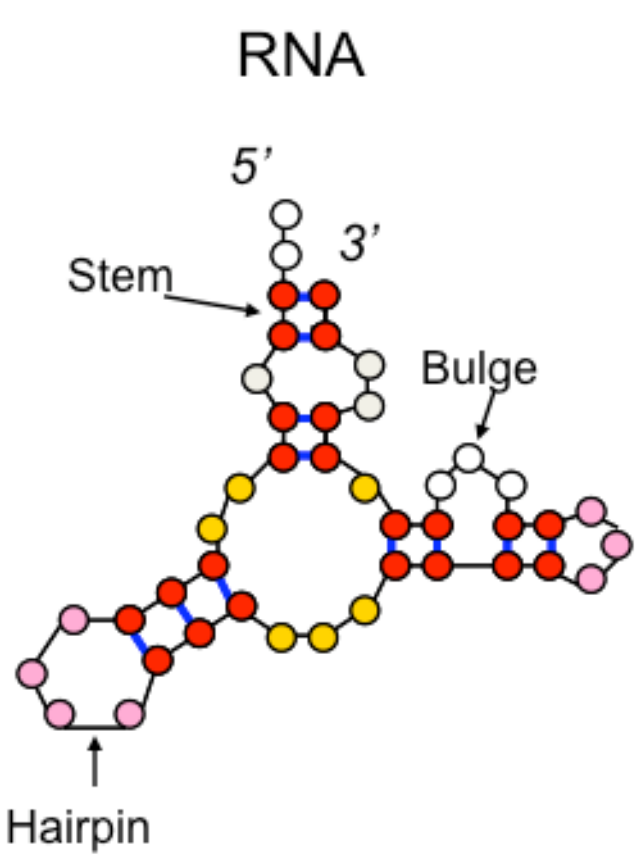
Outline

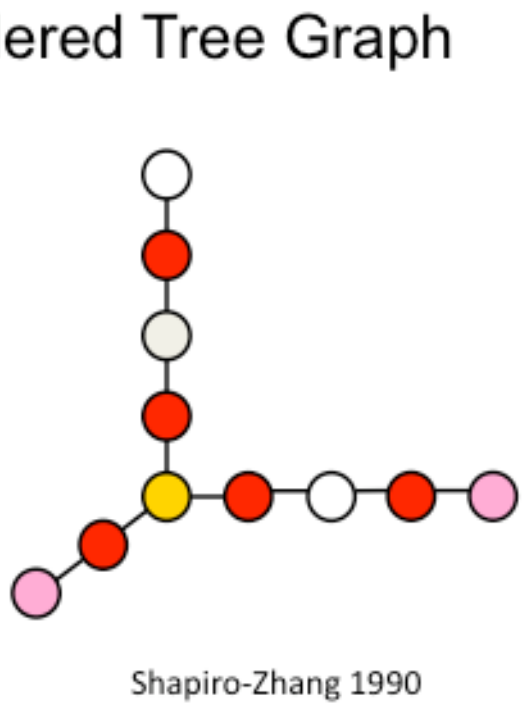
Representation of RNA Secondary Structures

- Sequences = Primary Structure
- Basis A, C, G, U can make pairings (hydrogen links), 4 levels of pairings :
  - *Watson-Crick* pairs : A—U and G — C
  - *Wobble* pairs (lower energy level) : G—U
  - pairs with very low level of energy : G—A or C—A
  - other pairs (rare) : actually any pair can occur. (Leontis N., Westhof E. 2001)

$\Rightarrow$ folding of the sequence in a secondary structure

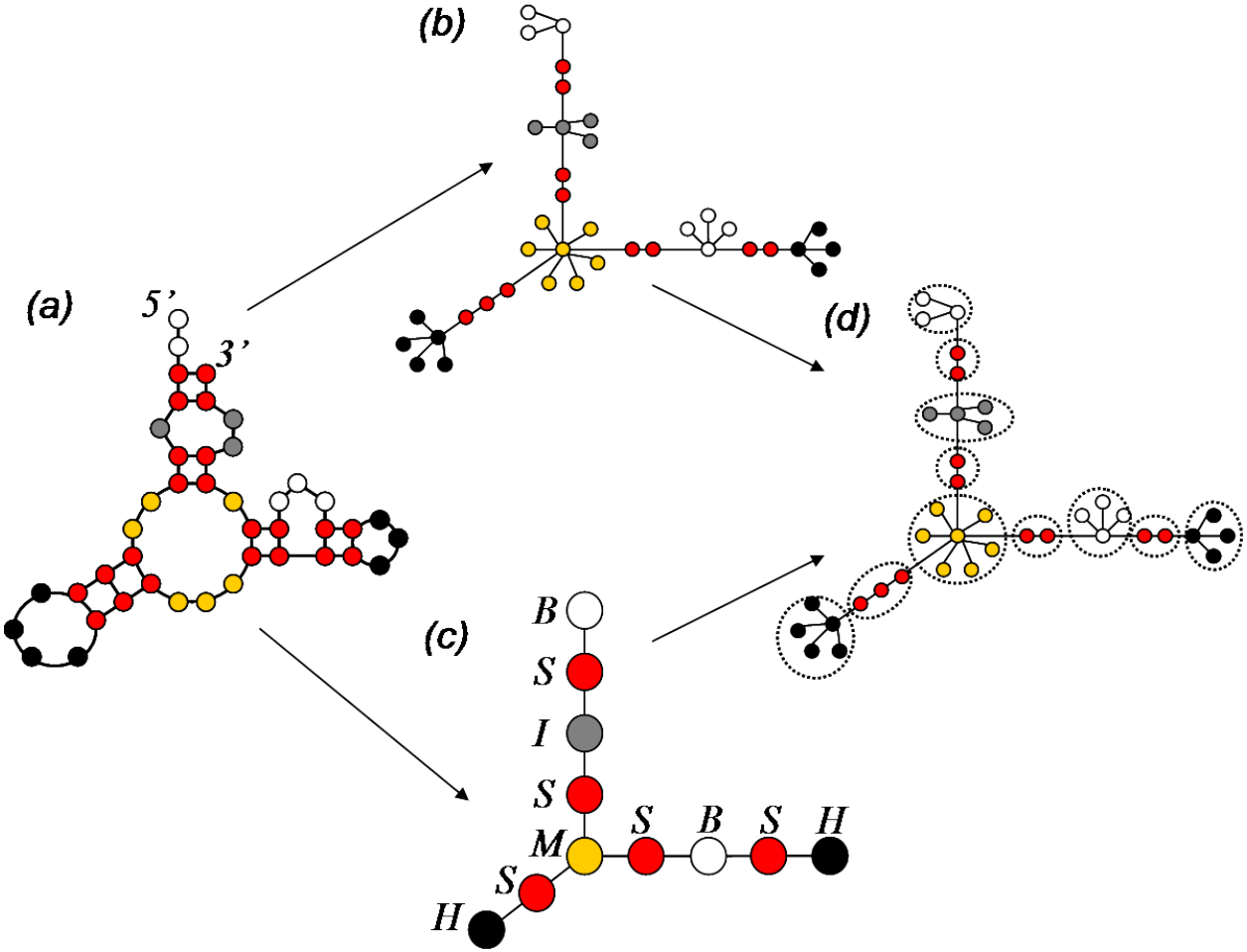
Representation of Secondary Structures of RNA





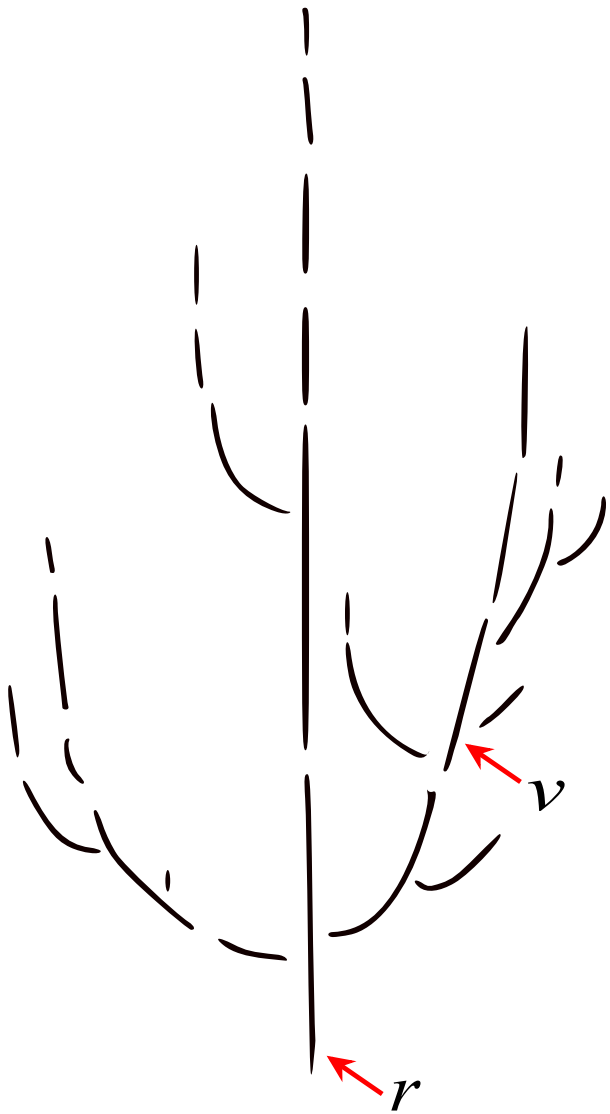
Toward a Multi-scale representation

- Ouangraoua et al. 2007

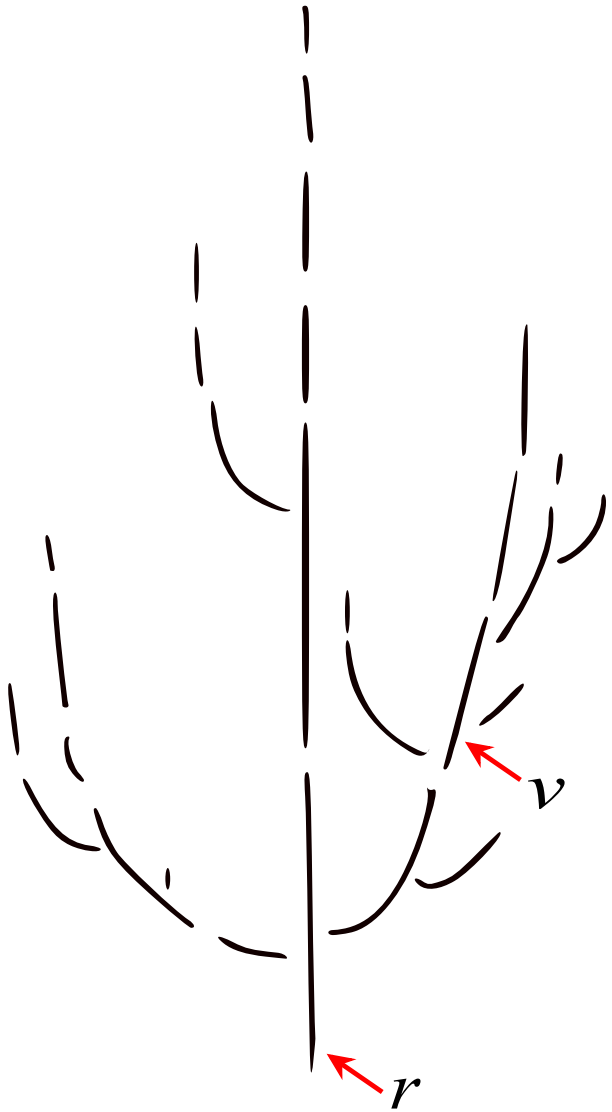


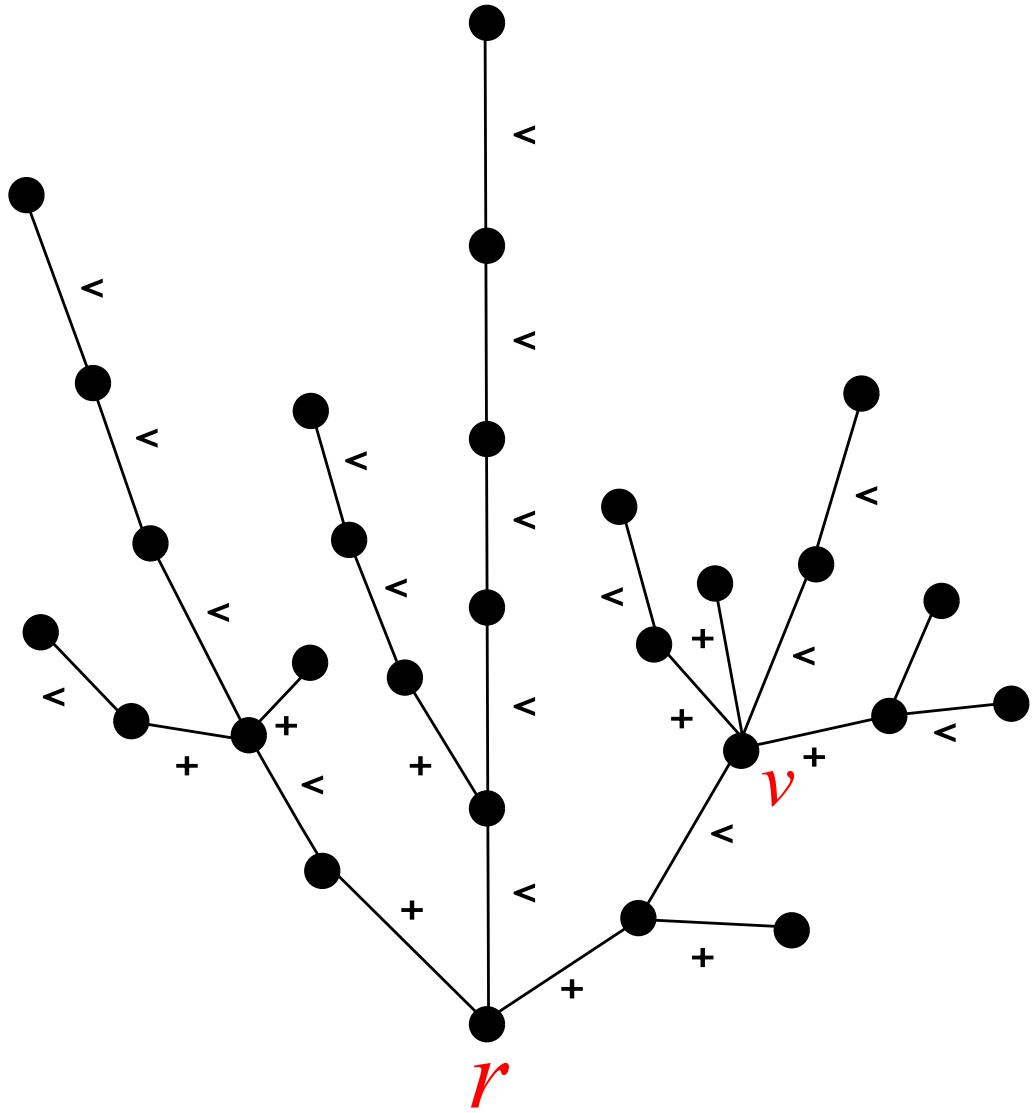
- A Multiple Graph Layers Model (Allali and Sagot , 2006)

Plant Architecture Modeling

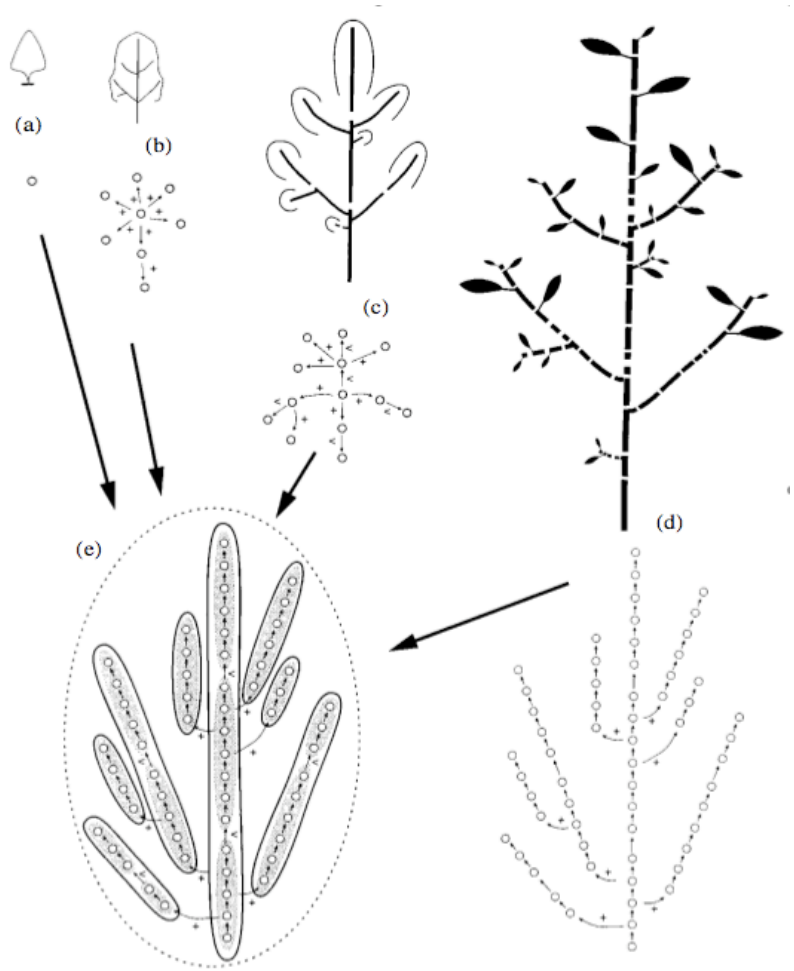


Plant Architecture Modeling



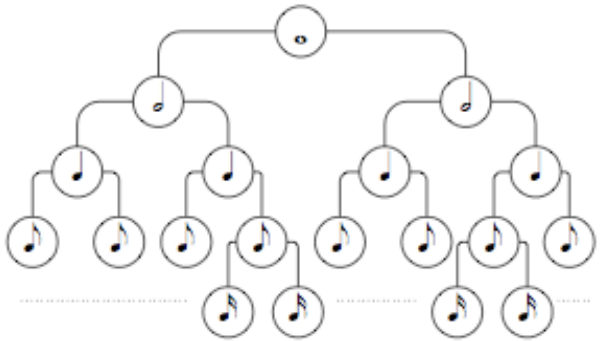


Plant Architecture Modeling



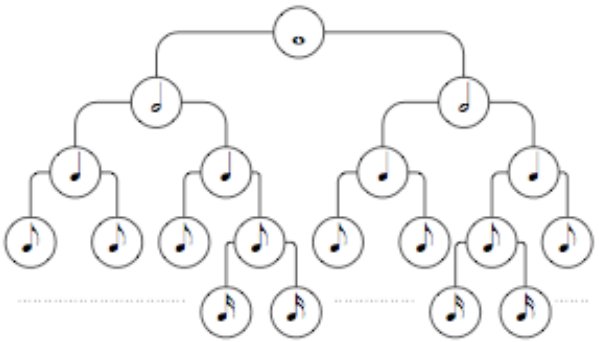
(Godin and Caraglio, 1998)

Tree Graph Representation of *Monophonic* Melody

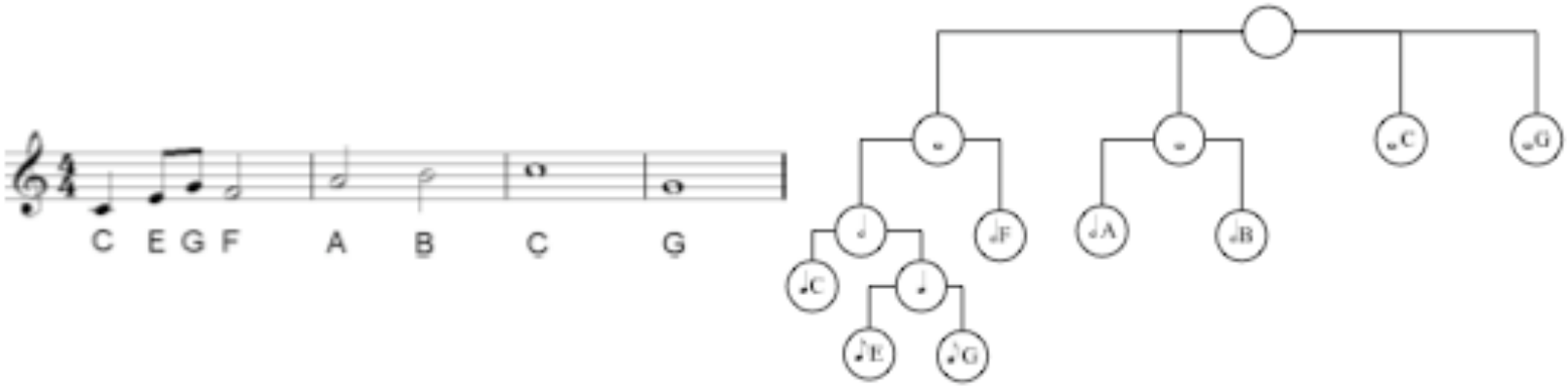


Hierarchy of note duration (Rizo et al., 2003)

Tree Graph Representation of *Monophonic* Melody

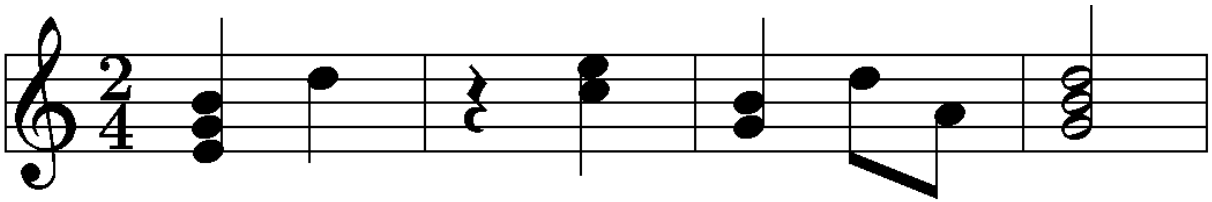


Hierarchy of note duration (Rizo et al., 2003)



Polyphony : Sequences of sequences

(Hanna and Ferraro, 2007)

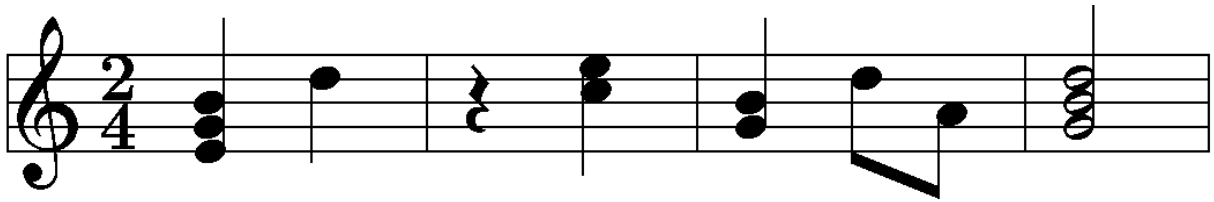


- Notes starting at the same time are grouped,
- Notes in a same chord are not ordered

- Problem with time overlapping : representation in linked notes

Polyphony : Sequences of sequences

(Hanna and Ferraro, 2007)

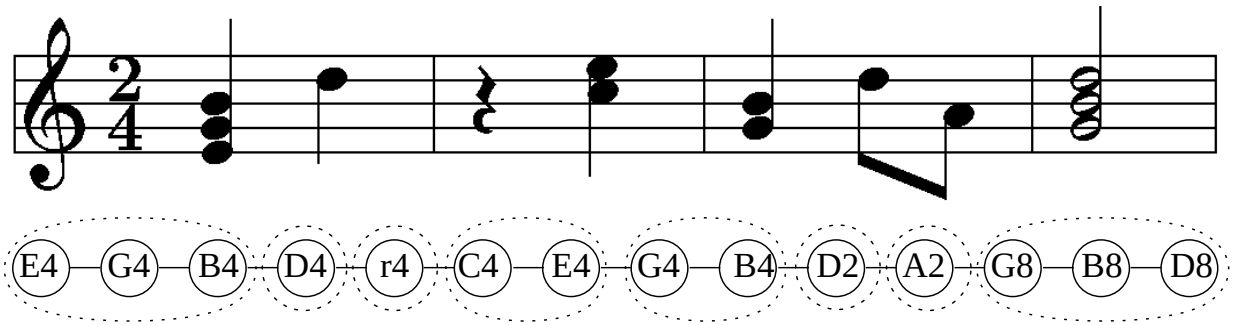


- Notes starting at the same time are grouped,
- Notes in a same chord are not ordered

- Problem with time overlapping : representation in linked notes

Polyphony : Sequences of sequences

(Hanna and Ferraro, 2007)

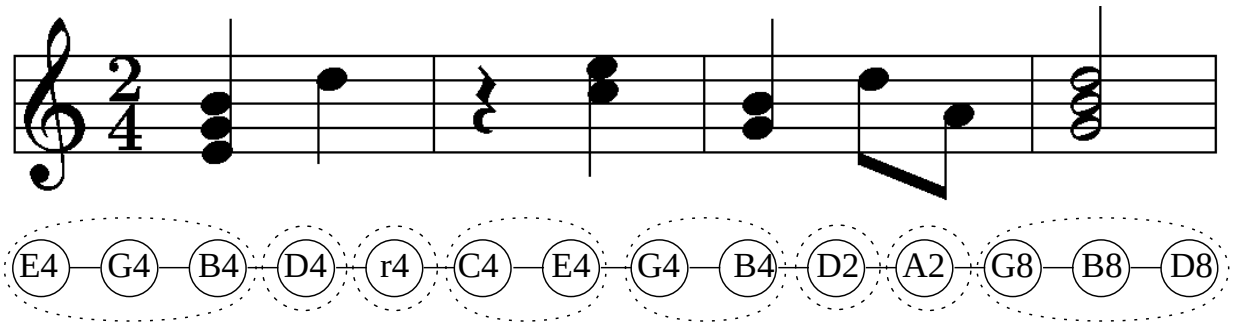


- Notes starting at the same time are grouped,
- Notes in a same chord are not ordered

- Problem with time overlapping : representation in linked notes

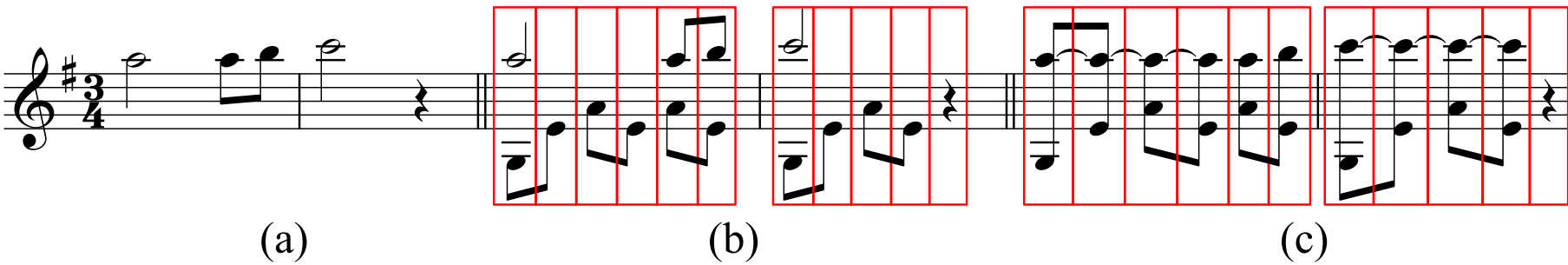
Polyphony : Sequences of sequences

(Hanna and Ferraro, 2007)



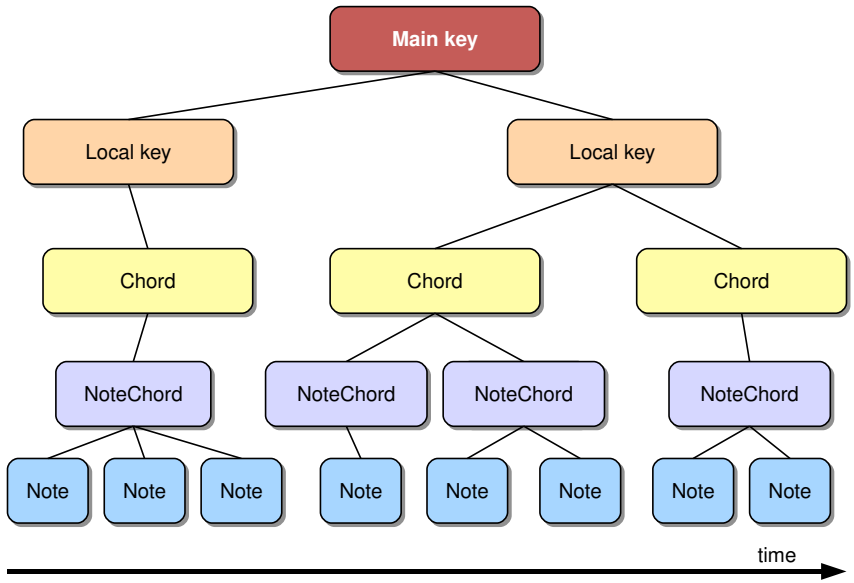
- Notes starting at the same time are grouped,
- Notes in a same chord are not ordered

- Problem with time overlapping : representation in linked notes

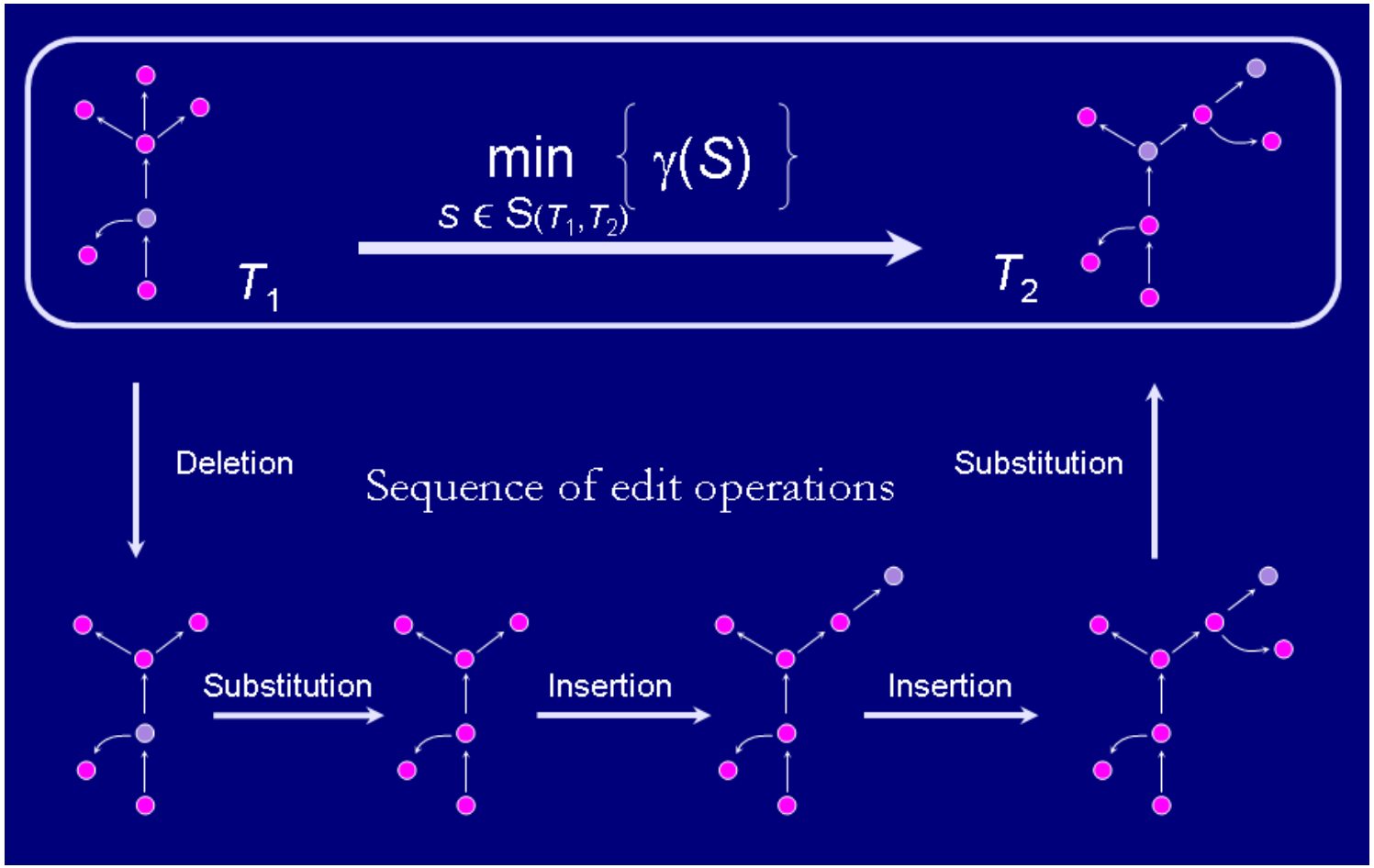


Western Music

- Main properties
  - Rhythm
  - **tonal Information**
- Different levels to be structured
- Tree Graph representation using 5 layers (Rocher, 2008)
  - Global tonality
  - Local tonality (modulations)
  - Chords (progression)
  - Groups of notes (homorhythmic)
  - Notes



Edit Distance between trees



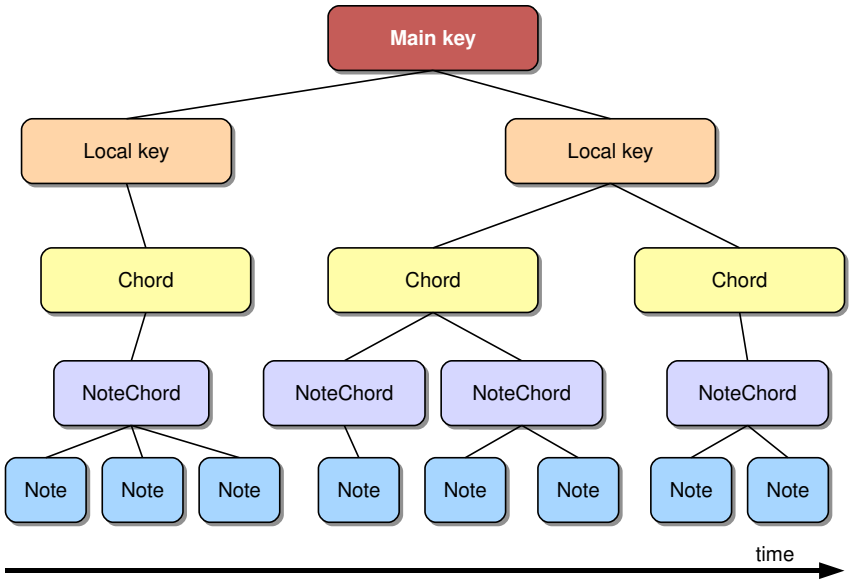
Several Variations

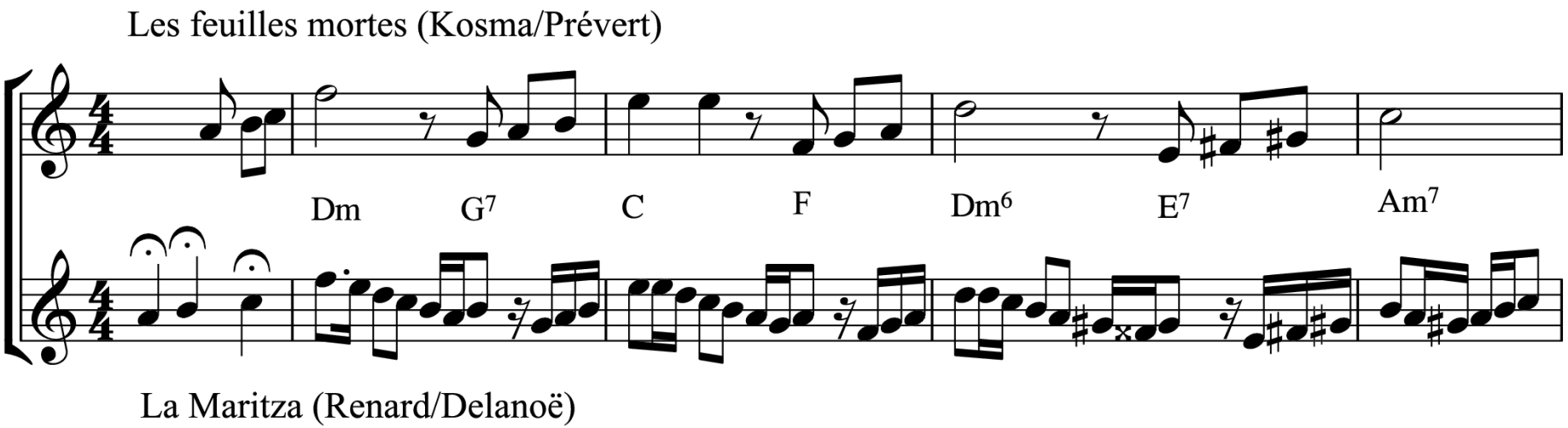
There is several methods based on tree edition principle:

- Constraints on tree height (Selkow, 1976)
- Ordered or Unordered trees (Zhang and Shasha, 1990, Zhang, 1996)
- Local edition (Ouangraoua et al., 2006)
- Alignment (Jiang et al., 1995, 2002)

Application to Plagiarism Detection

- Melodic similarity,
- Harmonic similarity,
- Combination of two.





First Experiments

Representation	Musical piece Similarity score		
R. Mack vs G. Harrison (1976)			
Query *Sweet Lord*			
	Sweet Lord	So Fine	*Essen* Rank 1
Note	143.8	14.0	20.0
Chord	35.7	**22.7**	20.5
Tree	179.6	**32.9**	30.0
Query *So Fine*			
	So Fine	Sweet Lord	*Essen* Rank 1
Note	153.2	14.0	21.8
Chord	32.5	**22.7**	21.8
Tree	187.7	**32.9**	25.3
Selle vs Gibb (1984)			
Query *Let It End*			
	Let It End	How Deep	*Essen* Rank 1
Note	159.0	**36.3**	30.2
Chord	35.7	14.5	33.5
Tree	194.7	**47.8**	37.3
Query *How Deep*			
	How Deep	Let It End	*Essen* Rank 1
Note	165.0	**36.3**	28.9
Chord	39.0	14.5	27.1
Tree	203.9	**47.8**	34.7

Outline

Musical and Algorithmic Perspectives

- Automatic detection of repetitions
  - Inference of musical structures (Allali et al., 2009) $\Rightarrow$ Verse - Chorus - Verse - Chorus
  - Longest repeated part (overlapping or not overlapping)
- no inference of musical structure. . . but comparison based on structural properties
  - Comparison of self-similarity matrices
  - Algorithmic problem: local alignment of 2D matrices
  - Musical applications: searching for music with a structural query

Examples:

- *Happiness is a Warm Gun*, Beatles
- *Paranoid Android*, Radiohead
- *Without you I'm Nothing*, Placebo

Long Term Perspectives

- Music recommendation systems
- Browsing music
- Pedagogy, musical games, evaluation of music performance
- Audio/score alignment, automatic accompaniment, synthesis
- Augmented listening (visualization of musical information)
- Active listening (musical properties)