Identify Homologous Words

Lydia

Bled 2011

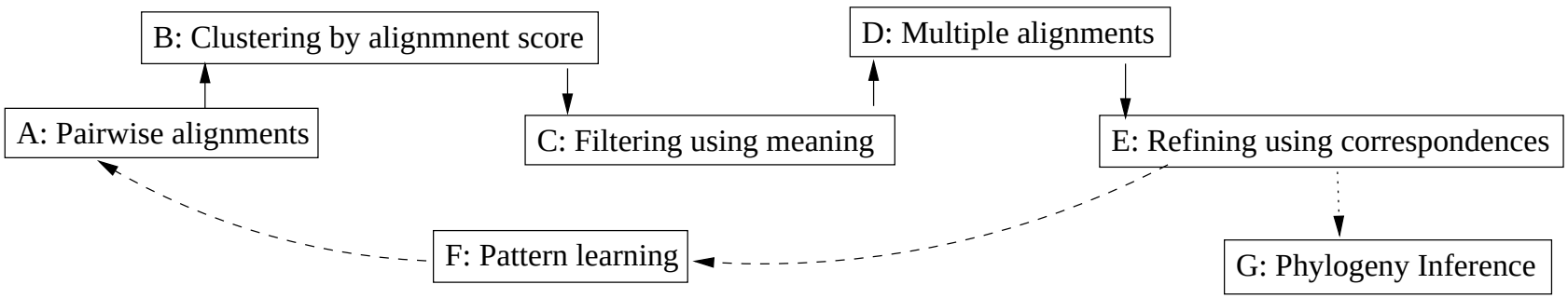
You may remind of ...

... my thesis project.

We show that

- bioinformatics methods works well on linguistics data
- finding homologous words is very similar to finding homologous proteins

Pipeline


B: Clustering by alignmnent score
D: Multiple alignments
A: Pairwise alignments
C: Filtering using meaning
E: Refining using correspondences
F: Pattern learning
G: Phylogeny Inference

Pipeline Input

- list of triples consisting of language, meaning and word
- encoded in utf-8
- words represent pronunciation not spelling
- words consists of at least one character
- one character consists of at least one sign

Pipeline Input

- list of triples consisting of language, meaning and word
- encoded in utf-8
- words represent pronunciation not spelling
- words consists of at least one character
- one character consists of at least one sign

The results are convincing but there are some groups of homologous words which seems to be wrong.

That's not right!

Mocovi	blood	l	e	w	o	ʔ
Wichi	blood	w	u	y	i	s
Wichi	vein,artery	w	u	y	i	s
Chorote	blood	w	o	y	i	s
Chorote	blood	y	o	y	i	s
Nivacle	blood	w	o	y	e	y
Wichi	vein,artery	n	o	y	i	h

That's not right!

Mocovi	blood	l	e	w	o	ʔ
Wichi	blood	w	u	y	i	s
Wichi	vein,artery	w	u	y	i	s
Chorote	blood	w	o	y	i	s
Chorote	blood	y	o	y	i	s
Nivacle	blood	w	o	y	e	y
Wichi	vein,artery	n	o	y	i	h

Making the threshold for being homologs (cognates) more restrictive destroy the complete set. :(

That's not right!

Mocovi	blood	l	e	w	o	?
Wichi	blood	w	u	y	i	s
Wichi	vein,artery	w	u	y	i	s
Chorote	blood	w	o	y	i	s
Chorote	blood	y	o	y	i	s
Nivacle	blood	w	o	y	e	y
Wichi	vein,artery	n	o	y	i	h

What's the problem?

What's the problem?

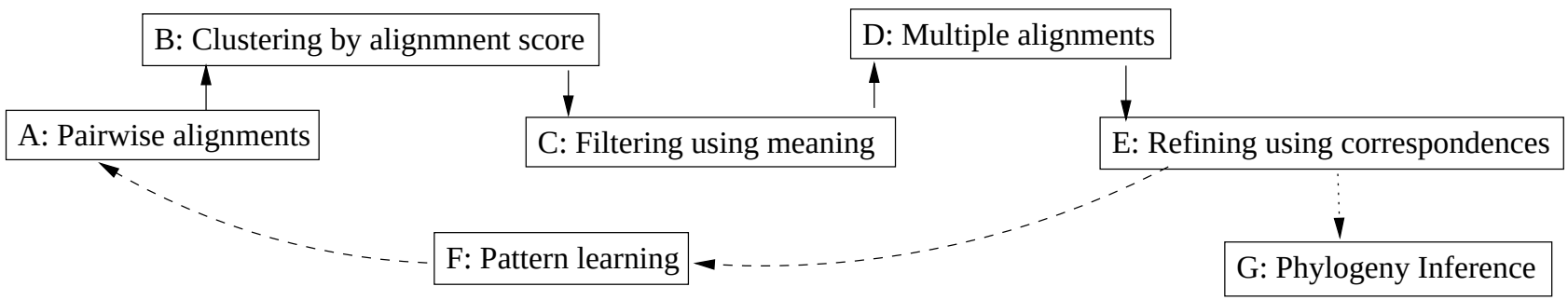
- alignment score for lewo? and wuyis could not exceed the threshold

What's the problem?

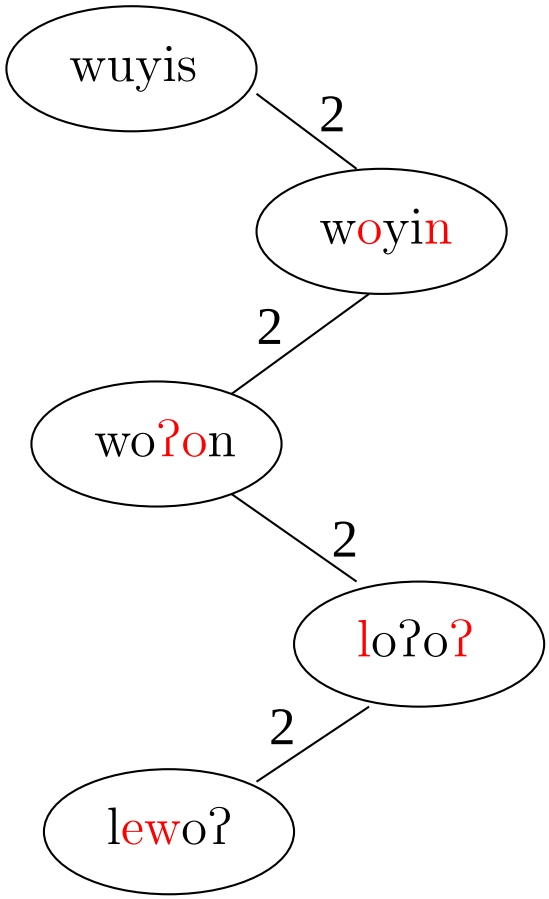
- alignment score for lewo? and wuyis could not exceed the threshold
- there must be path between lewo? and wuyis in the graph representing the cognate/homolog relation

What's the problem?

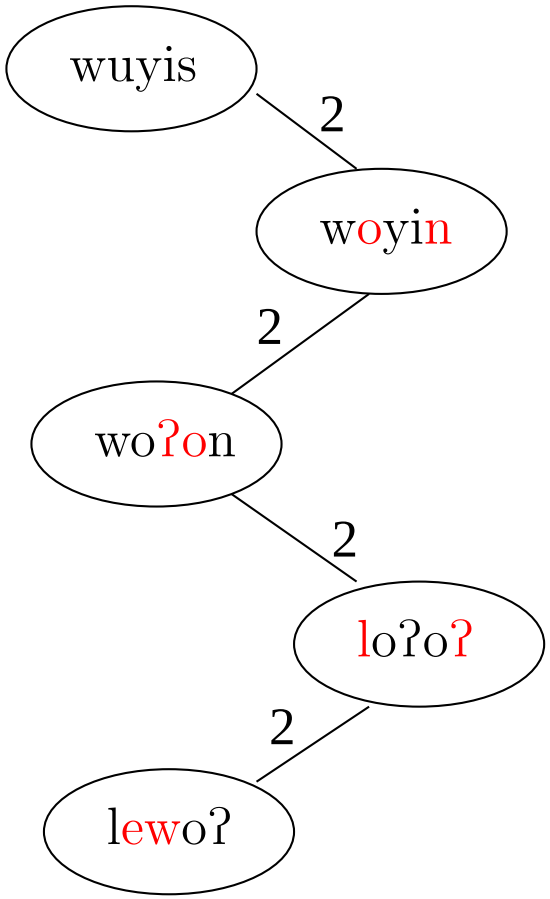
- alignment score for lewo? and wuyis could not exceed the threshold
- there must be path between lewo? and wuyis in the graph representing the cognate/homolog relation
- either in step C or in step E the linking words of the path are removed from the cognate set



What's the problem?

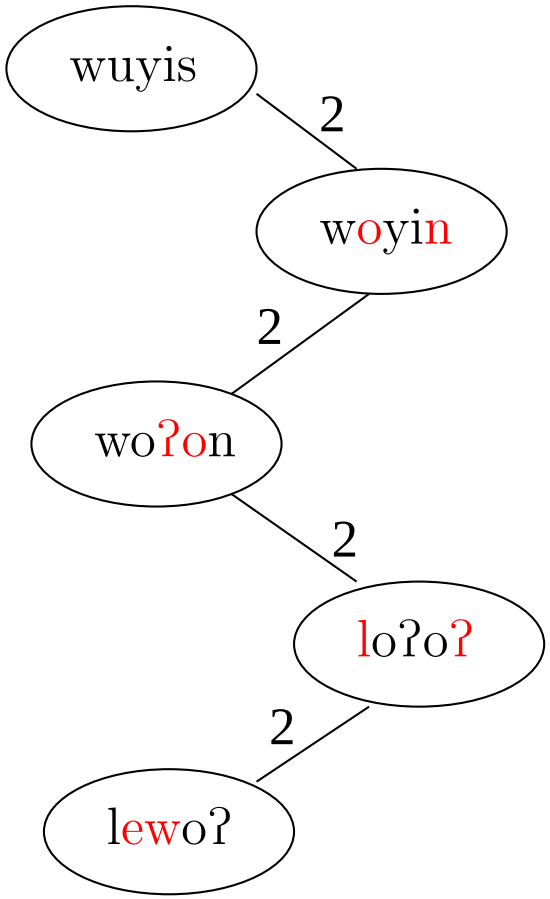


What's the problem?



wuyis	blood
woyin	seek, look for
woʔon	marry
loʔoʔ	prostitute
lewoʔ	blood

What's the problem?



wuyis	blood
woyin	seek, look for
woʔon	marry
loʔoʔ	prostitute
lewoʔ	blood

Solution

We need

a better clustering algorithm. It should ensure that

1. resulting groups are highly connected
2. the cut should removes as few as possible edges

Solution - Spectral Clustering

Steps

1. remove tree-like structures

Solution - Spectral Clustering

Steps

1. remove tree-like structures
2. construct the Graph Laplacian $L = D - A$

Solution - Spectral Clustering

Steps

1. remove tree-like structures
2. construct the Graph Laplacian $L = D - A$
3. calculate the second eigenvalue (algebraic connectivity a) and corresponding eigenvector (fiedler vector $\vec{v}$)

Solution - Spectral Clustering

Steps

1. remove tree-like structures
2. construct the Graph Laplacian $L = D - A$
3. calculate the second eigenvalue (algebraic connectivity a) and corresponding eigenvector (fiedler vector $\vec{v}$) [In case of questions referring this step, contact Peter]
4. normalize the algebraic connectivity by the number of nodes ($\bar{a}$)

Solution - Spectral Clustering

Steps

1. remove tree-like structures
2. construct the Graph Laplacian $L = D - A$
3. calculate the second eigenvalue (algebraic connectivity a) and corresponding eigenvector (fiedler vector $\vec{v}$) [In case of questions referring this step, contact Peter]
4. normalize the algebraic connectivity by the number of nodes ($\bar{a}$)
5. if $\bar{a} < T$ (T - predefined threshold), split graph using $\vec{v}$ and return to Step 1

What about "lewo?"

It works!

1. "lewo?" is removed from the cognate set
2. it is now grouped together with "lawo?" (occurred in several languages, means family)
3. we could make the threshold for being cognate less restrictive such that we could better resolve the relation in the short words

Thanks to

Peter, Michael and you