Never miss a profile

String-based search using EMPOP2 and application to phylogenetic alignment

Alexander Röck[1] Arne Dür[1]
Stefan Troger[2] Martin Pircher[2] Walther Parson[2]

[1] Institute of Mathematics, University of Innsbruck, Austria

[2] Institute of Legal Medicine, Innsbruck Medical University, Austria

Haploid DNA markers in forensic genetics
Berlin 22-24 April 2010

Motivation

Standardization of alignment of some mtDNA profiles is difficult

- $\Rightarrow$ ISFG rules [Carracedo et al. FSI 2000] not sufficient
- $\Rightarrow$ recommendations by [Wilson et al. FSI 2002] lead to artificial mutations when applied rigorously
- $\Rightarrow$ phylogenetic approach by [Bandelt & Parson IJLM 2008] incorporates current knowledge of phylogeny (real mutations)
- $\Rightarrow$ active area of scientific discussion (e.g. [Den Hartog et al. FSI:GSS 2009])

Phylogenetic vs. rule-based alignment

Current situation

- mtDNA profiles are reported as differences to rCRS
- forensic mtDNA databases store profiles as differences to rCRS
- searches within databases mostly rely on the annotation of profiles
- as new data are generated rule-based as well as phylogenetic approach may require further expansion

Phylogenetic vs. rule-based alignment

Current situation

- mtDNA profiles are reported as differences to rCRS
- forensic mtDNA databases store profiles as differences to rCRS
- searches within databases mostly rely on the annotation of profiles
- as new data are generated rule-based as well as phylogenetic approach may require further expansion

How to overcome these annotation difficulties in forensic mtDNA databases?

Phylogenetic vs. rule-based alignment

Current situation

- mtDNA profiles are reported as differences to rCRS
- forensic mtDNA databases store profiles as differences to rCRS
- searches within databases mostly rely on the annotation of profiles
- as new data are generated rule-based as well as phylogenetic approach may require further expansion

How to overcome these annotation difficulties in forensic mtDNA databases?

EMPOP2 allows for an alignment independent search!

EMPOP2 string search

Input

query profile as differences to rCRS or in FASTA-like format

Search

query profile as well as database profiles are converted to FASTA-like format and then compared to each other

Output

transcript = “how to convert a neighbouring database profile into the query profile”

2 main options

- pattern match vs. literal match
- possibility to disregard indels in length variants at positions 16188, 16193, 309, and 455

Phylogenetic vs. formal rules

Example 1 – USA0600976 [Diegoli et al. FSI:G 2009]

Phylogenetic alignment

16172C 16183C
16188T 16189C 16193.1C
16223T 16320T 16519C
73G 150T 195C 263G
309.1C 315.1C

Formal rules

16172C 16183DEL
16193.1C 16193.2C
16223T 16320T 16519C
73G 150T 195C 263G
309.1C 315.1C

The resulting FASTA-like strings are equal.

Comparison of search results of EMPOP2

Search in EMPOP2 (database size = 7330 full CR profiles)

number of differences to query profile	number of hits in EMPOP2 position-based phylogenetic alignment	position-based rule-based alignment	string-based	haplo-group
0	1	0	1	L3e2b
1	4	0	4	L3e2b
2	6	3	6	L3e2b
3	3	12	7	
4	5	6	6	
5	6	7	6	
6+	7305	7302	7300	

Numerous possibilities to align a sequence

Example 1 – USA0600976 [Diegoli et al. FSI:G 2009] – phylogenetic alignment

16172C 16183C 16188T 16189C 16193.1C 16223T 16320T 16519C 73G 150T 195C 263G 309.1C 315.1C

Other variants

- 16183DEL 16193.1C 16193.2C
- 16183C 16187.1T 16189C
- 16183C 16188T 16188.1C 16189C
- 16183C 16188DEL 16193.1C 16193.2C
- 16182.1C 16183C 16187T 16189C

Tolerating alignments with 13+2 differences to rCRS leads to 798 possibilities.

Understanding the output of EMPOP2

Example 1 – USA0600976 [Diegoli et al. FSI:G 2009] – rule-based alignment

16172C 16183- 16193.1C 16193.2C 16223T 16320T 16519C 73G 150T 195C 263G 309.1C 315.1C

Nearest database profile (0 differences, haplogroup L3e2b)

16172C 16183C 16188T 16189C 16193.1C 16223T 16320T 16519C 73G 150T 195C 263G 309.1C 315.1C

Transcript from database to query profile

no changes necessary

Understanding the output of EMPOP2

Example 1 – USA0600976 [Diegoli et al. FSI:G 2009] – rule-based alignment

16172C 16183- 16193.1C 16193.2C 16223T 16320T 16519C 73G 150T 195C 263G 309.1C 315.1C

Nearest database profile (0 differences, haplogroup L3e2b)

16172C 16183C 16188T 16189C 16193.1C 16223T 16320T 16519C 73G 150T 195C 263G 309.1C 315.1C

Transcript from database to query profile

no changes necessary

Database profile + transcript = phylogenetic alignment of query profile (haplogroup L3e2b)

16172C 16183C 16188T 16189C 16193.1C 16223T 16320T 16519C 73G 150T 195C 263G 309.1C 315.1C

Understanding the output of EMPOP2

Example 2 – IG1089 [Nohira et al. IJLM 2010]

16150T 16183C 16185T 16189C 16193DEL 16217C 16234T 16519C 73G 151T 197G 263G 315.1C 523DEL 524DEL 546G 573.1C

Search in EMPOP2 (database size = 7330 full CR profiles)

number of differences to query profile	number of hits in EMPOP2	haplogroup
0	0	
1	0	
2	0	
3	0	
4	0	
5	2	B4d3
6+	7328	

Understanding the output of EMPOP2

Example 2 – IG1089 [Nohira et al. IJLM 2010]

16150T 16183- 16185.1T 16189DEL 16217C 16234T 16519C
73G 151T 197G 263G 315.1C 523DEL 524DEL 546G 573.1C

Nearest database profile (5 differences, haplogroup B4d3)

16183C 16185T 16186Y 16189C 16217C 16234T 16519C
73G 151T 152C 197G 263G 309.1C 315.1C 546G

Transcript from database to query profile

C16150T C152T A523DEL C524DEL -573.1C
C16193DEL C309.1DEL (ignored)

Understanding the output of EMPOP2

Example 2 – IG1089 [Nohira et al. IJLM 2010]

16150T 16183- 16185.1T 16189DEL 16217C 16234T 16519C
73G 151T 197G 263G 315.1C 523DEL 524DEL 546G 573.1C

Nearest database profile (5 differences, haplogroup B4d3)

16183C 16185T 16186Y 16189C 16217C 16234T 16519C
73G 151T 152C 197G 263G 309.1C 315.1C 546G

Transcript from database to query profile

C16150T C152T A523DEL C524DEL -573.1C
C16193DEL C309.1DEL (ignored)

Database profile + transcript = phylogenetic alignment of query profile (haplogroup B4d3)

16150T 16183C 16185T 16189C 16193DEL 16217C 16234T 16519C
73G 151T 197G 263G 315.1C 523DEL 524DEL 546G 573.1C

Conclusion

Inconsistent alignment may lead to underestimation of the frequency of the input profile.

Using the search engine of EMPOP2 relieves the forensic user of the burden of alignment issues and guarantees that matching sequences are found.

EMPOP2 string search

- input can be differences to rCRS or FASTA-like string
- possibility to ignore indels that are forensically not relevant
- differences from database to query profile given in output

References & Acknowledgements

Carracedo A et al
DNA Commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing.
Forensic Sci Int ***110***(2) 2000

Wilson A et al
Recommendations for consistent treatment of length variants in the human mtDNA control region.
Forensic Sci Int ***129***(1) 2002

Bandelt & Parson
Consistent treatment of length variants in the human mtDNA control region: a reappraisal
IJLM ***122***(1) 2008

Den Hartog et al
The impact of jumping alignments on mtDNA population analysis and database searching
FSI:GSS ***2*** 2009

This work is funded by the FWF–Austrian Science Fund through Translational Research project L397 “EMPOP–an innovative human mtDNA database”.