Bioinformatika és genomanalízis az orvostudományban

Korreláció, klaszterezés

Cserző Miklós

2020

https://semmelweis.zoom.us/j/96102872458?pwd=Rk1PL2tqS21sdlUwc3B4eDFCZkNKQT09

A mai előadás

- Adatsorok korrelációja
- A klaszterező algoritmusok típusai
- A klaszterezésben használt metrikák
- Hierarchikus klaszterezés
- K-közép klaszterezés
- Klaszterezés a biológiában
- Klaszterező alkalmazások

Mi a klaszterezés

- Adathalamzok belső szerkezetének elemzése, feltárása
- Ez alapján az adatok csoportosítása tulajdonságaik szerint
- Vagy hierarchikus rendszerbe szervezése
- A ‘klaszterezés’ egy általános cél, amit sok módon lehet elérni
- A módszerek nyelvhasználata jelentősen eltér

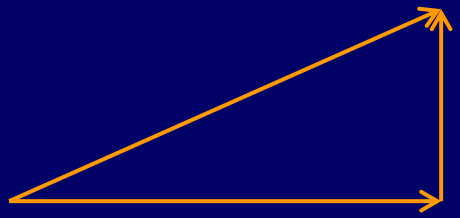
Hogyan mérjük a hasonlóságot

- Az egyes adatpontokra jellemző tulajdonság – metrika
- Ezeket hasonlítjuk össze egymással
- Az összehasonlítás a klaszterezés alapja: távolságmátrix
- A metrika lehet:
  - Közvetlenül megfigyelhető fizikai mennyiség
  - Absztrakt mennyiség

Metrikák vektortérben

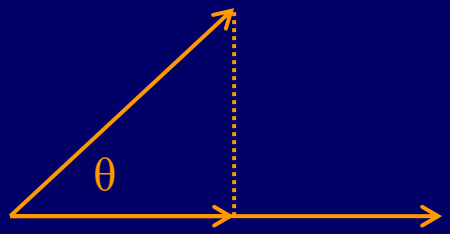
Távolság:

$$d^2 = (a_1 - b_1)^2 + (a_2 - b_2)^2 + .. + (a_n - b_n)^2$$



Skalár szorzat:

$$p = a_1 \bullet b_1 + a_2 \bullet b_2 + a_3 \bullet b_3 + .. + a_n \bullet b_n$$

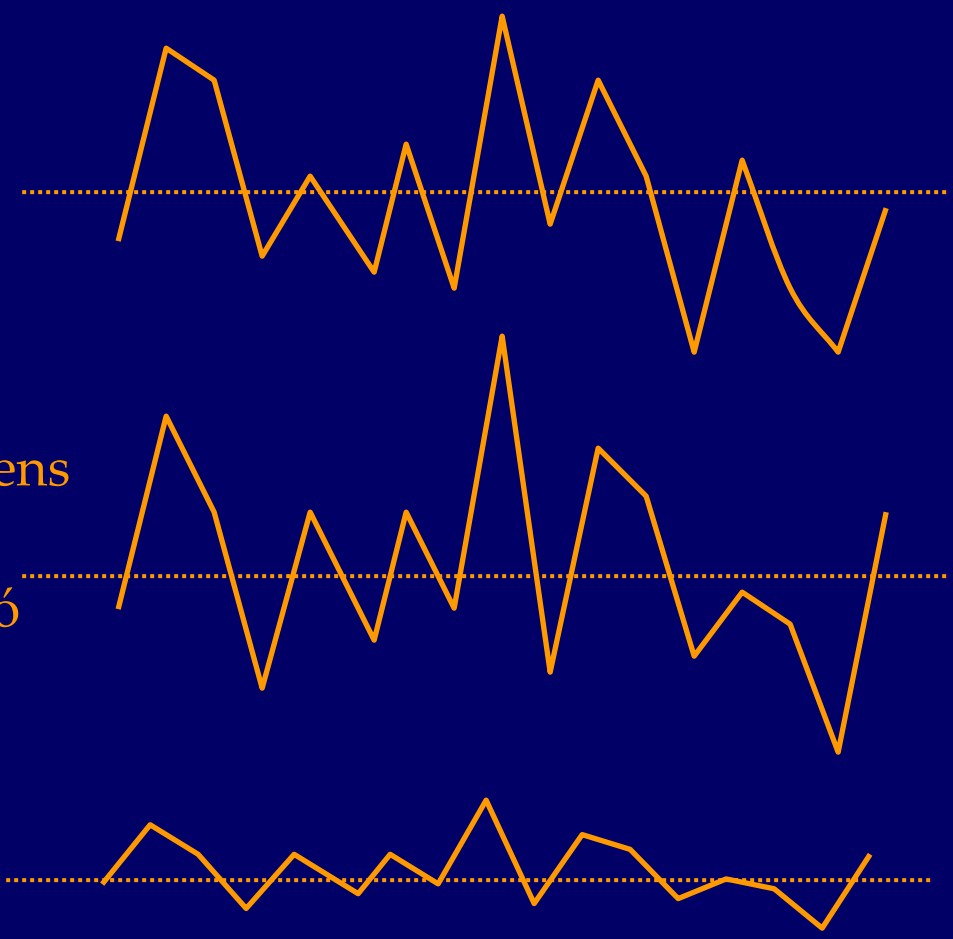


$$p = |a| \bullet |b| \cos(\Theta)$$

Adatsorok korrelációja

$$r_{xy} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\bullet\left(y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2 \bullet \sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2}}$$

- x_i, y_i: adatpontok
- X, Y: adatok átlaga
- A korrelációs koefficiens -1 és +1 közé esik
- -1: teljes antikorreláció
- +1: teljes korreláció
- 0: véletlen eset
- A szignifikancia nő a pontsor hosszával



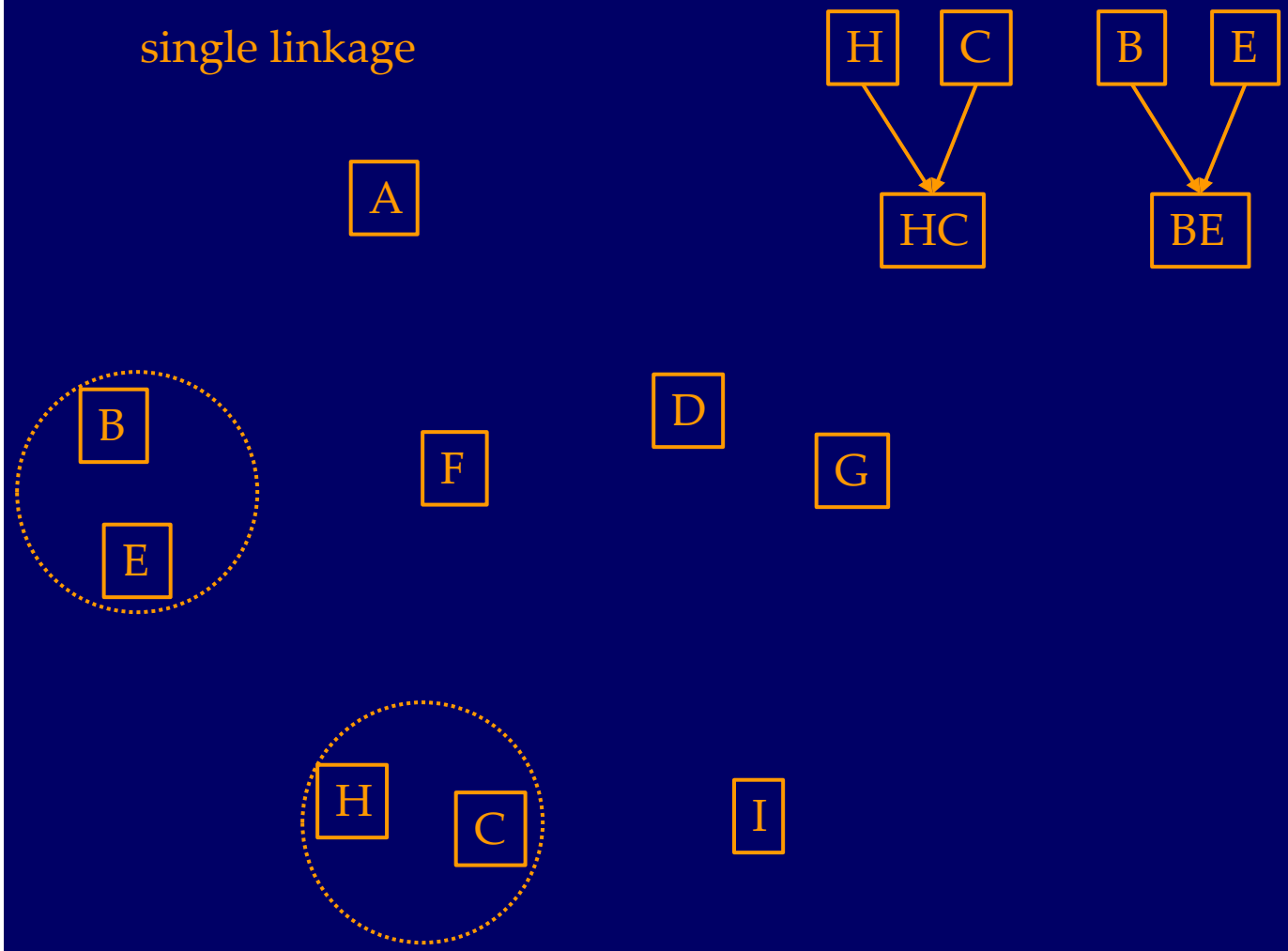
Hierarchikus klaszterezés

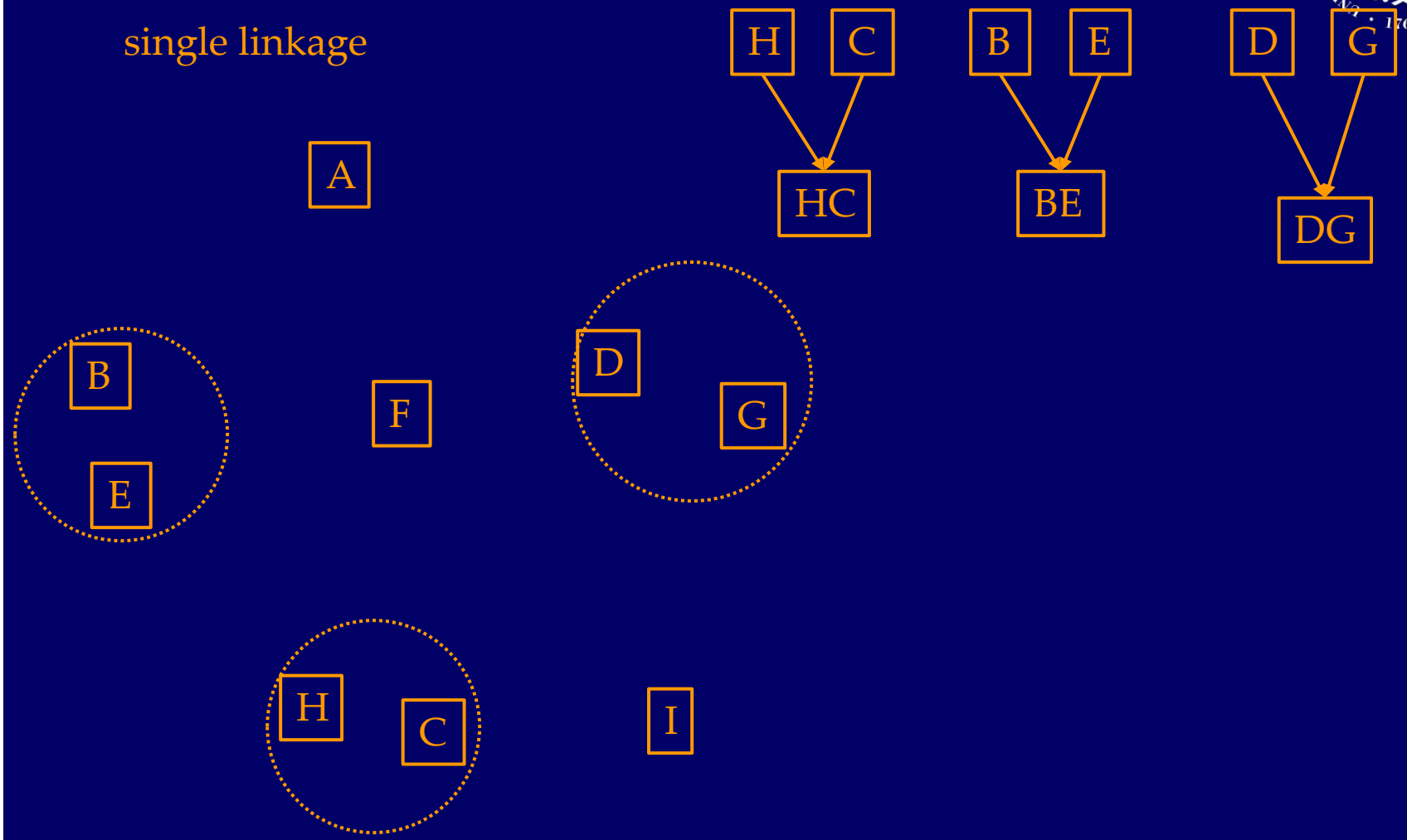
- Az alap: a hasonló elemek közel esnek egymáshoz
- Az eredmény: dendogram (‘fa’ elrendezés)
- A probléma komplexitása: n^3
- Nagy adathalmazok esetén lassú
- Agglomeratív eljárás: minden adatpont külön klaszter
- Az eljárás során egyesítjük a klasztereket

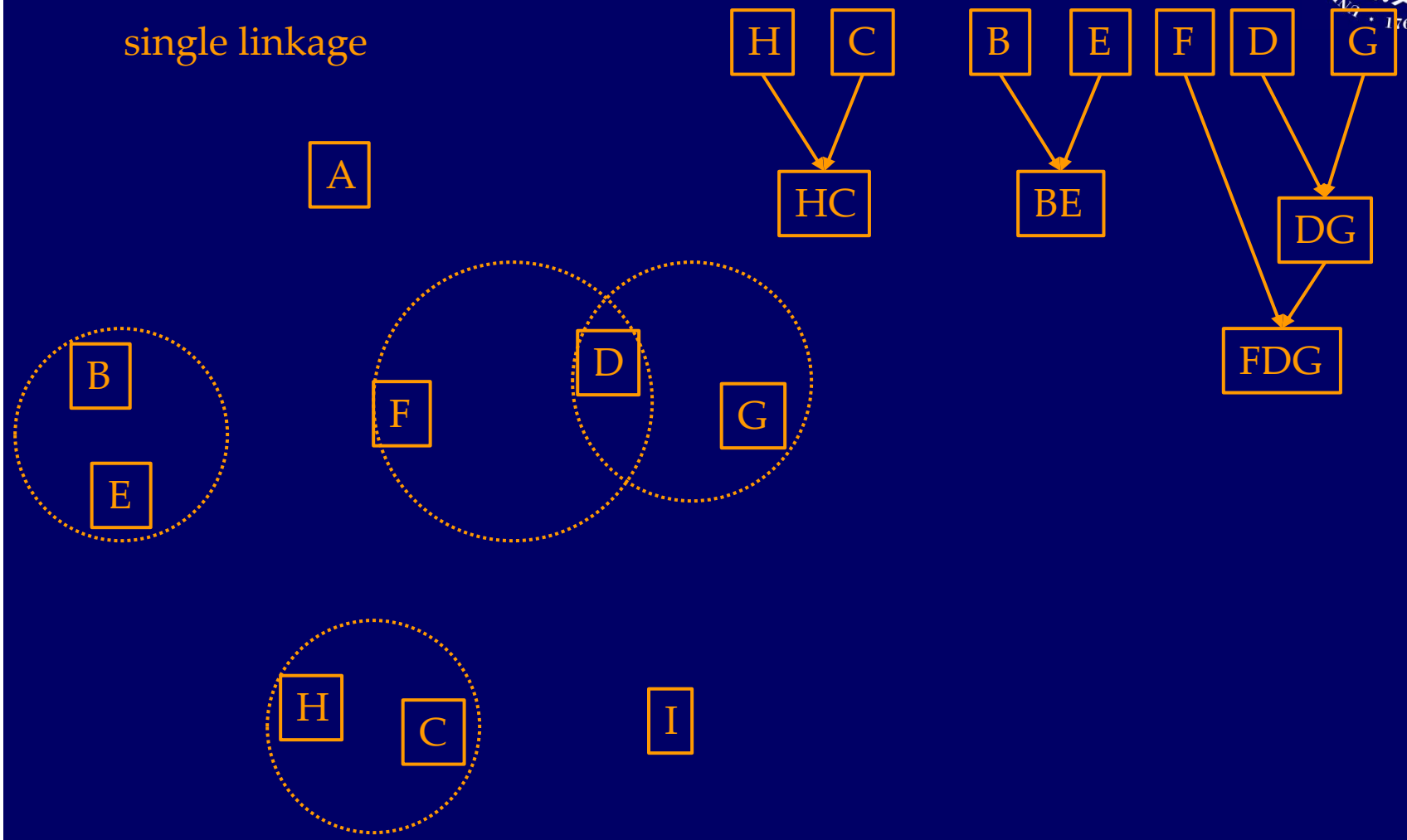
Lehetséges klaszterező szabályok

- ‘single linkage’ – a klaszterek elemeinek minimális távolságát keresi
- ‘complete linkage’ – az elemek maximális távolságát vizsgálja
- ‘UPGMA’ – a két részklaszter elemeinek átlagos távolságát nézi
- Stb…

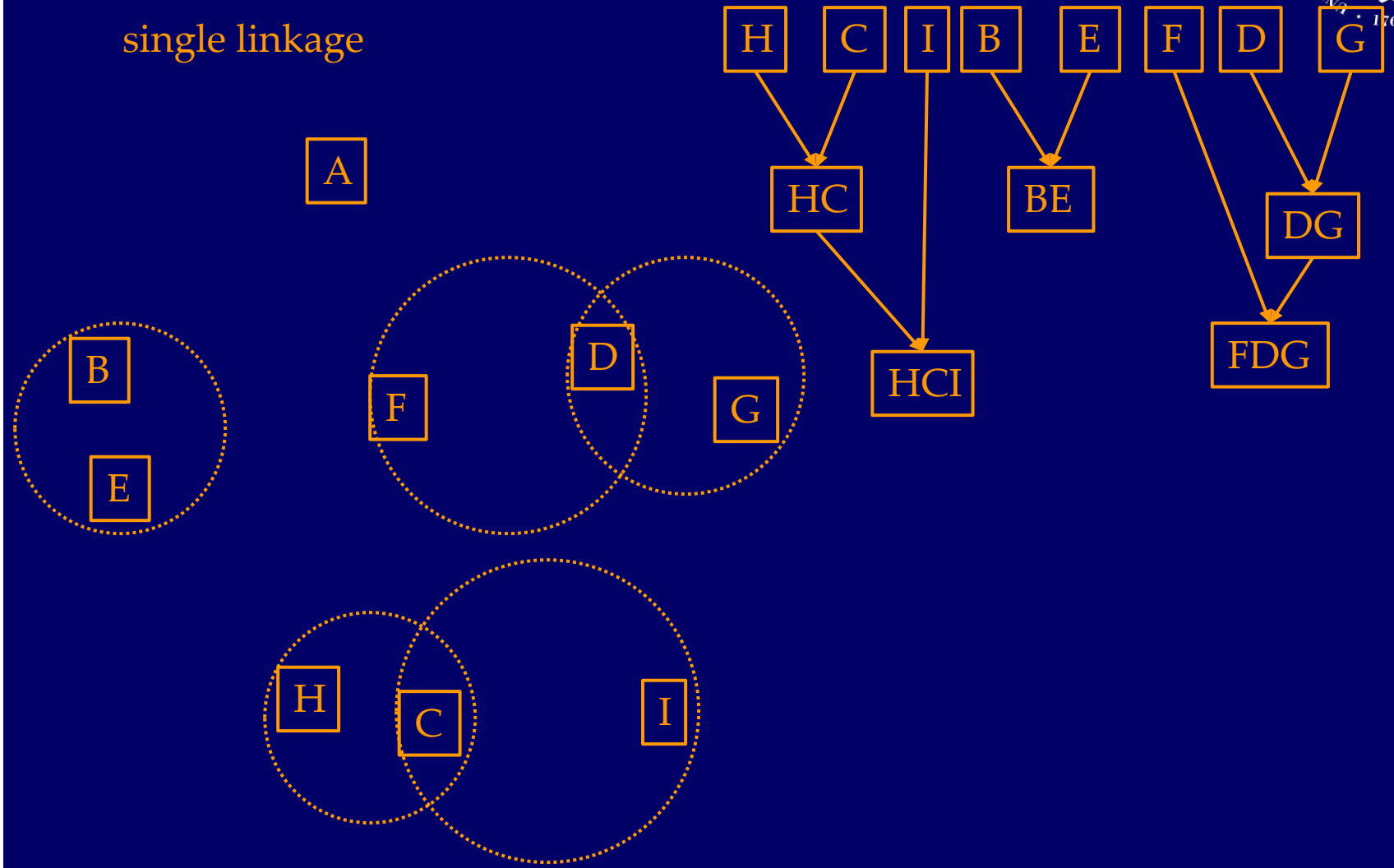
single linkage



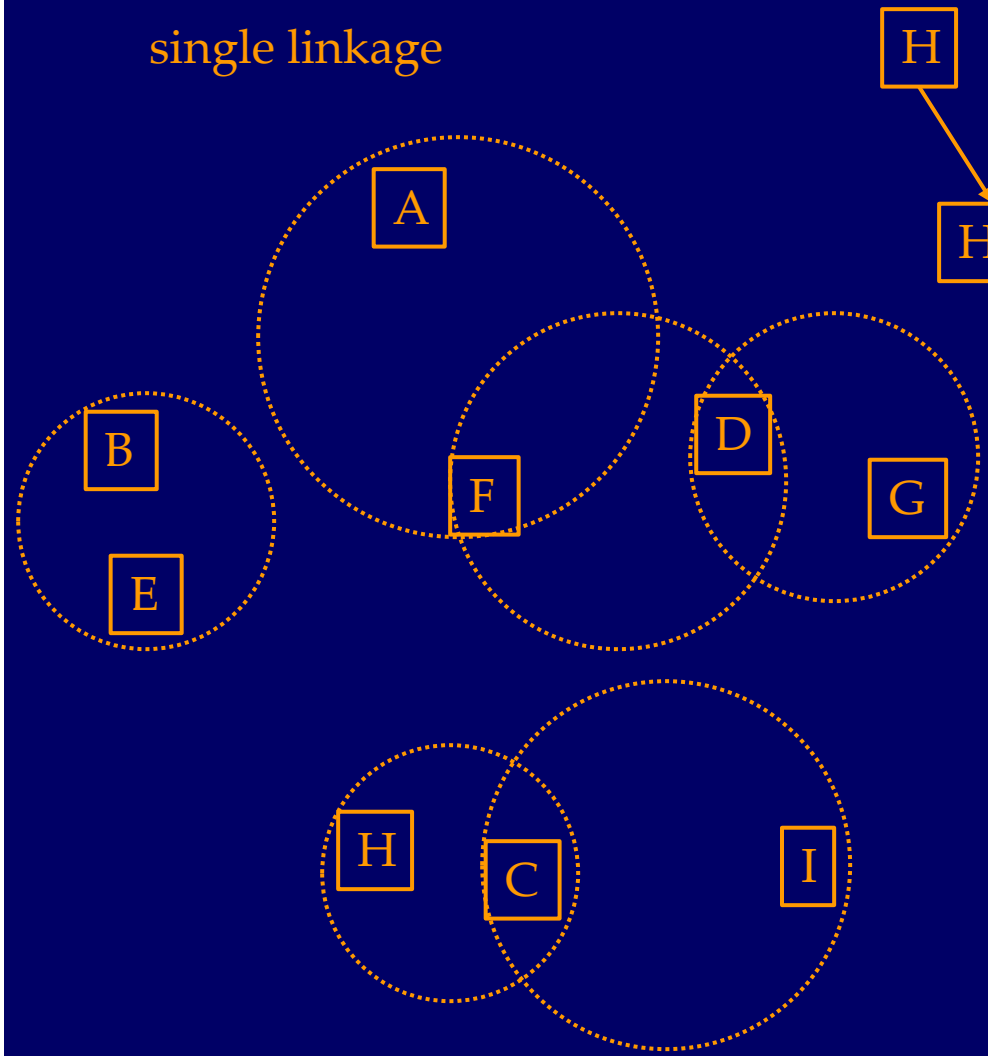

single linkage
A
B
E
F
D
G
H
C
I
H
C
HC
B
E
BE
D
G
DG

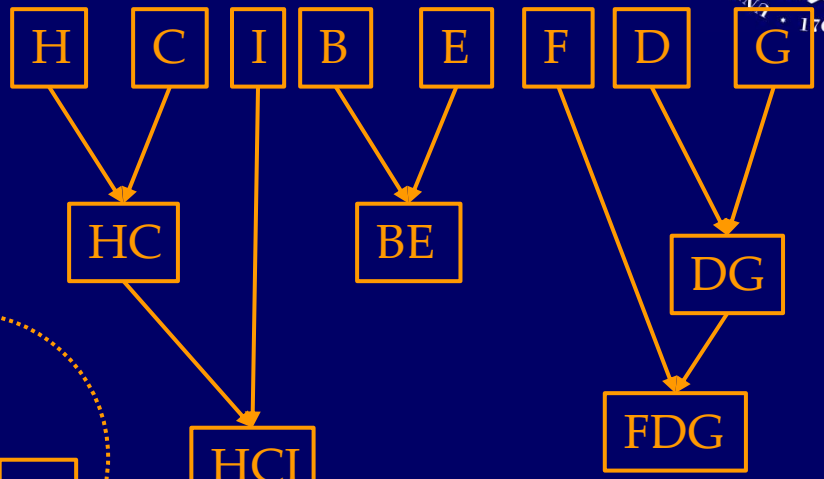

single linkage
A
B
E
F
D
G
H
C
I
H
C
HC
B
E
BE
F
D
G
DG
FDG

single linkage

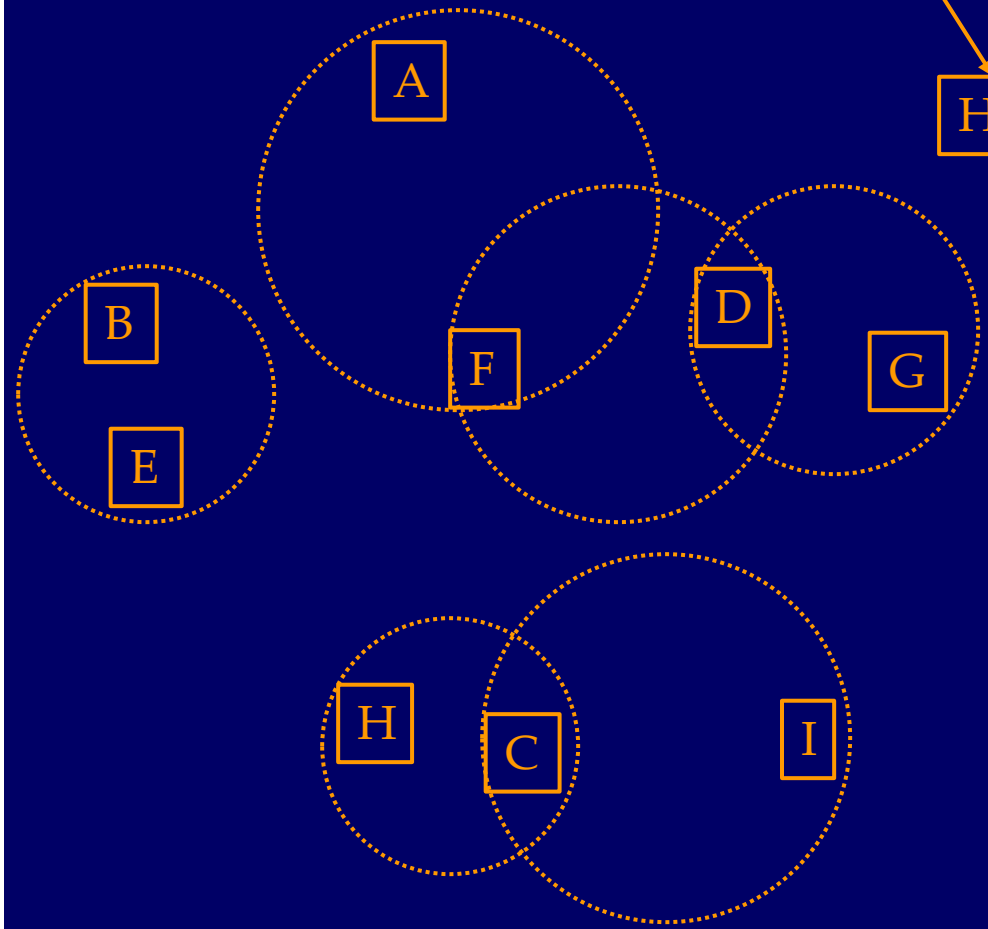


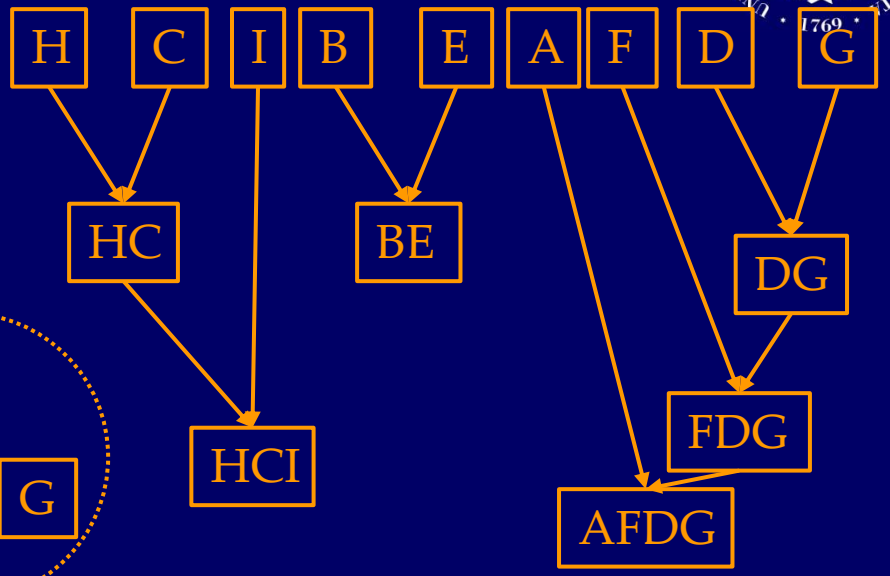
single linkage



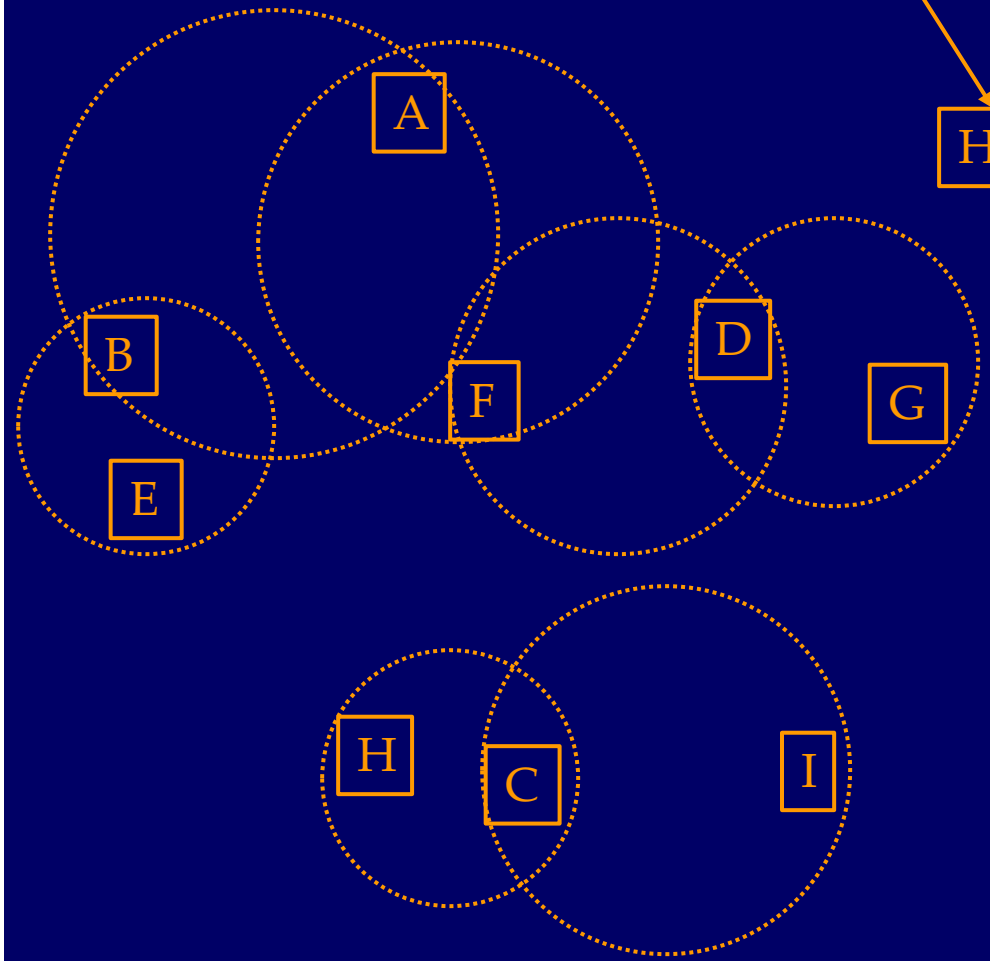


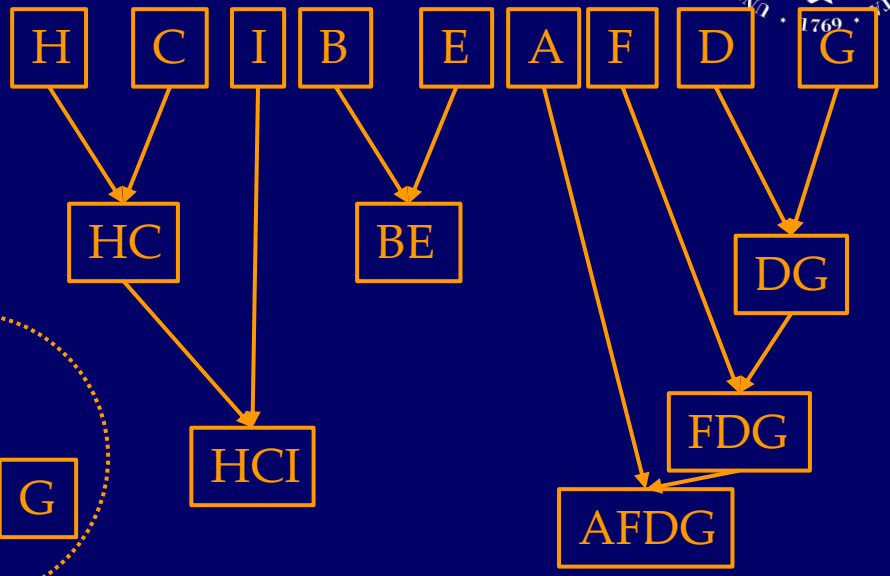
single linkage

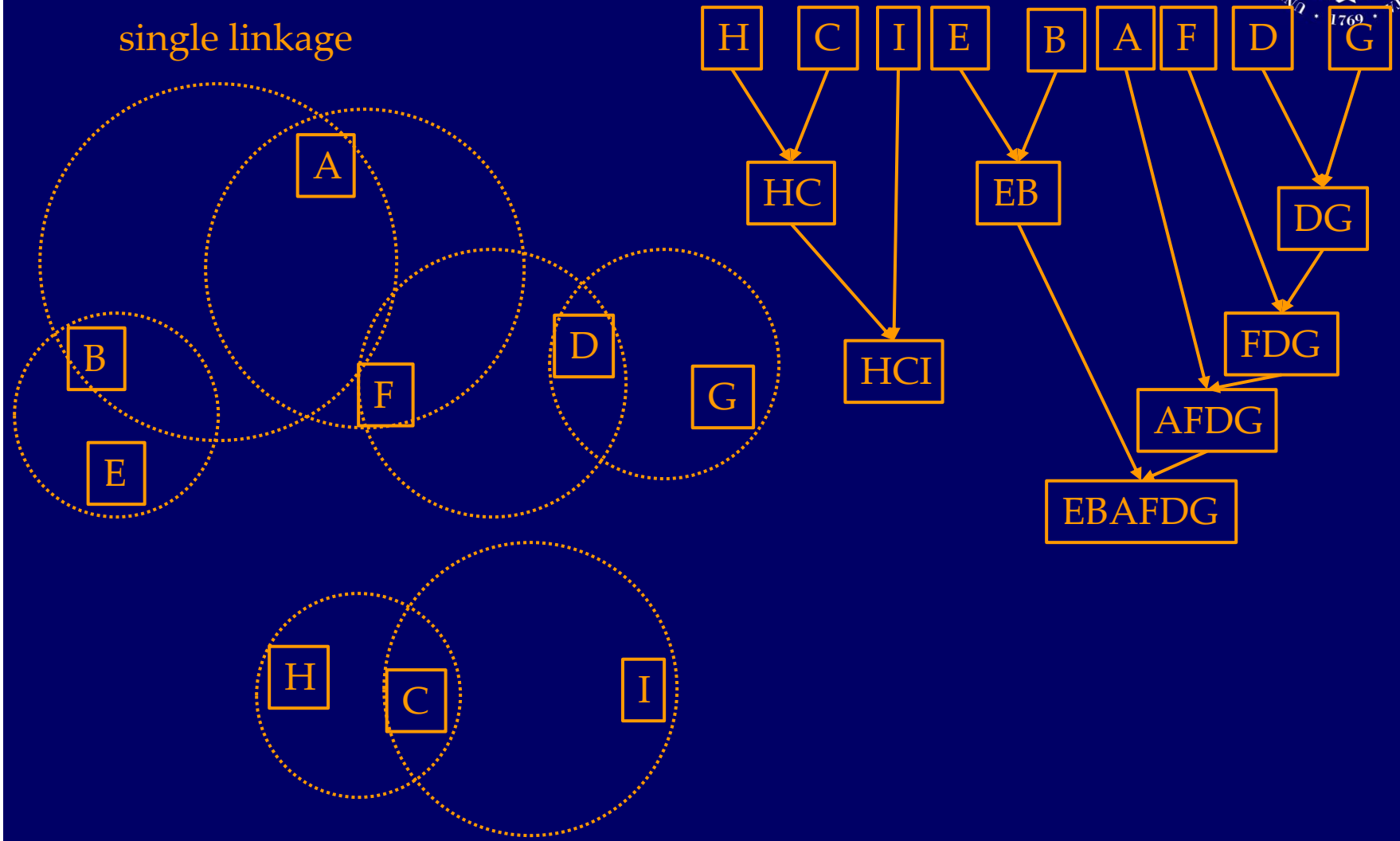




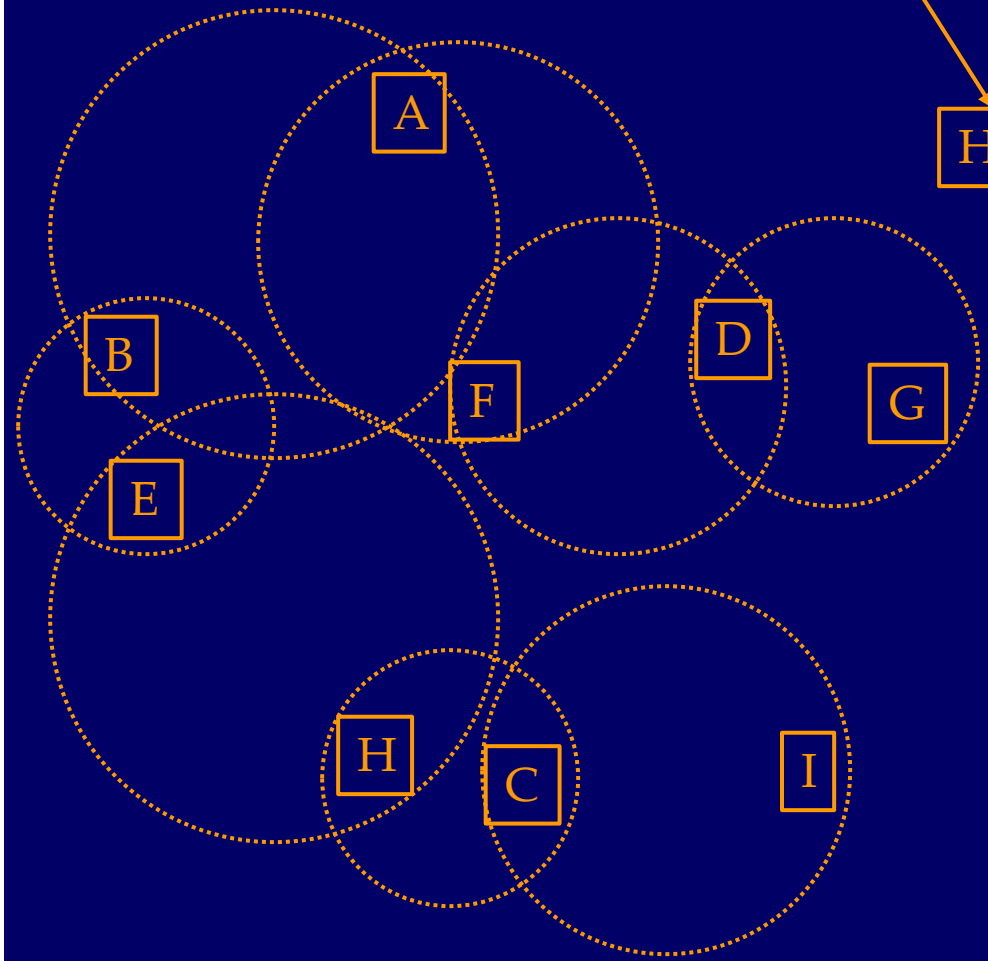
single linkage

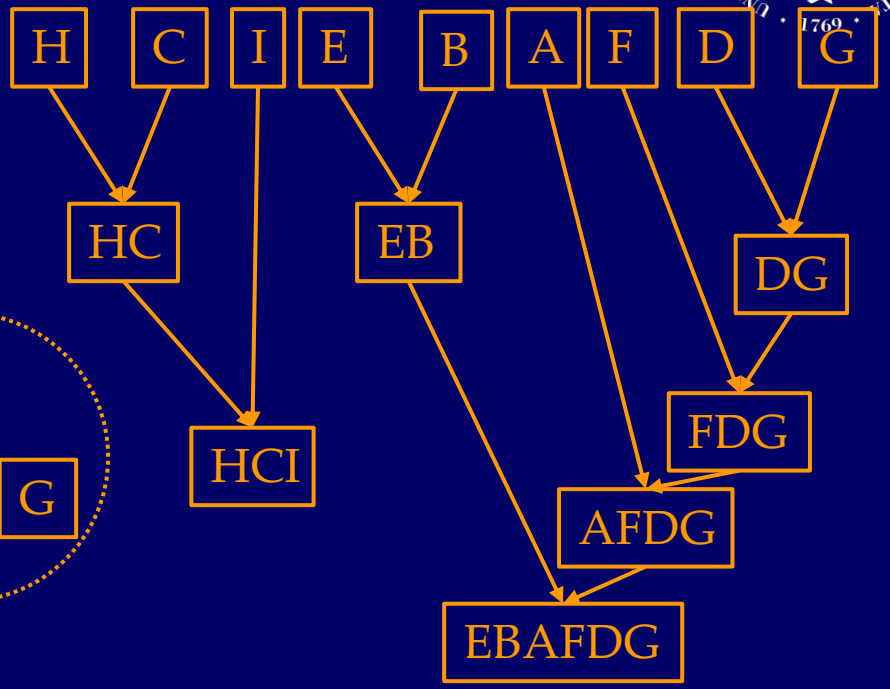




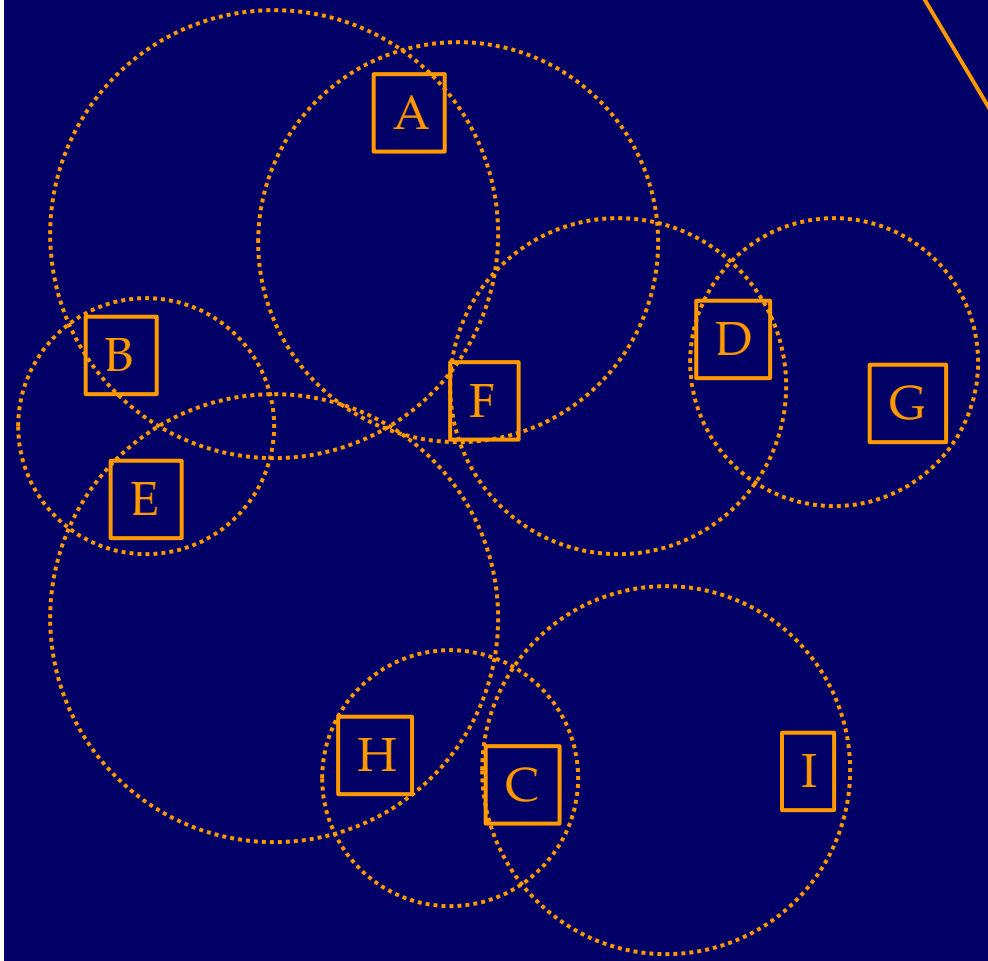

single linkage
A
B
E
F
D
G
H
C
I
H C I E B A F D G
HC
EB
DG
HCI
FDG
AFDG
EBAFDG

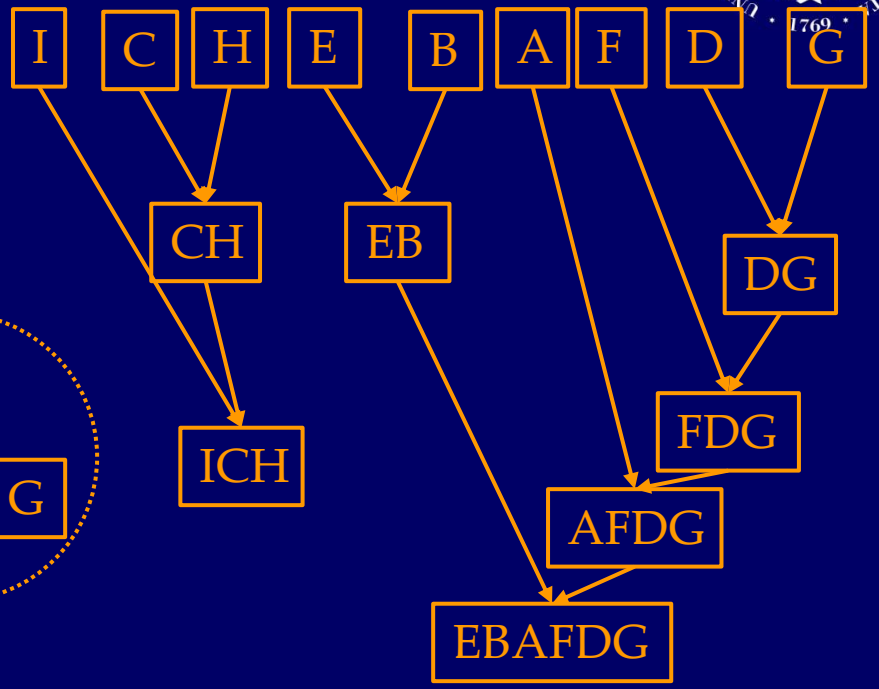
single linkage

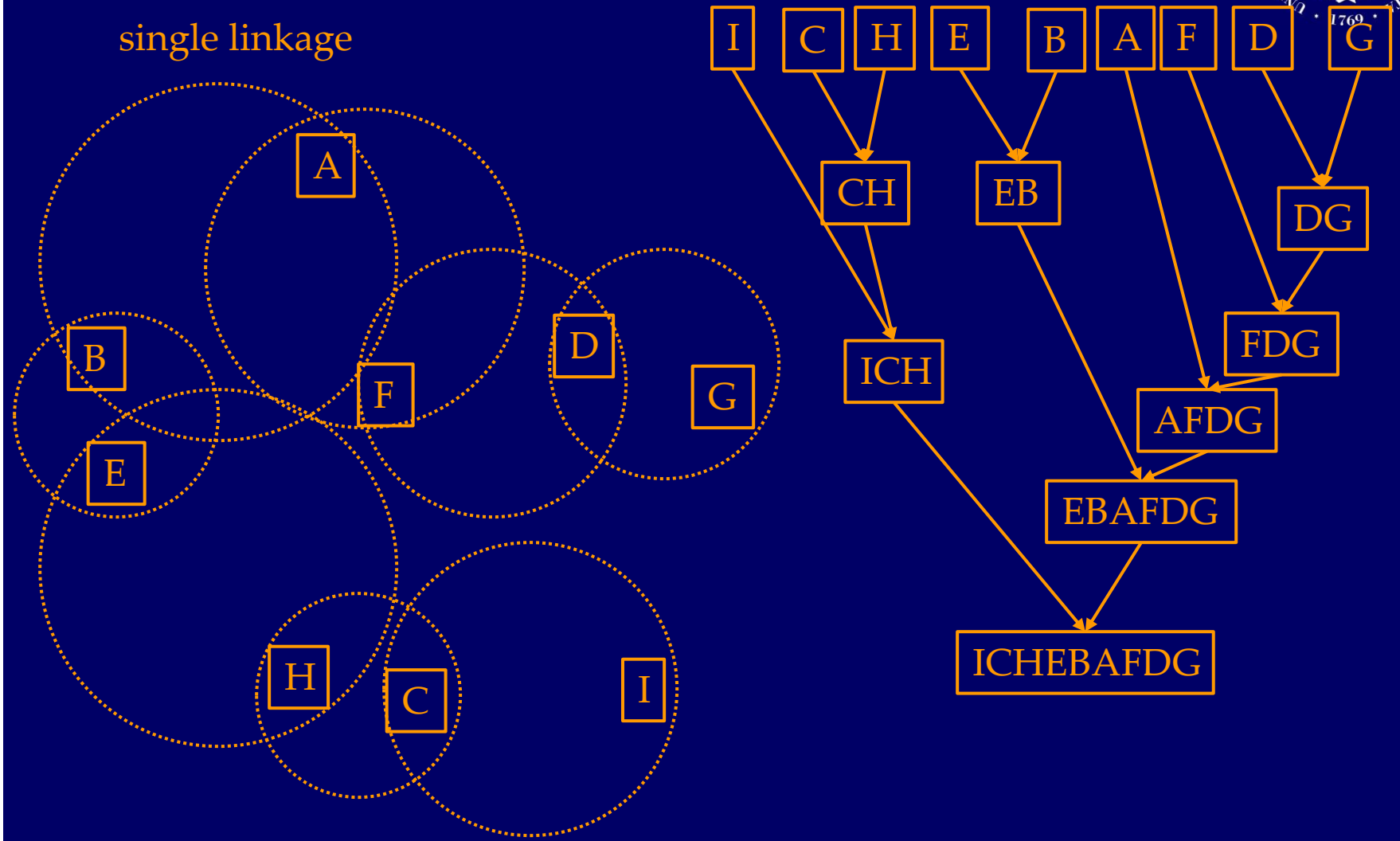




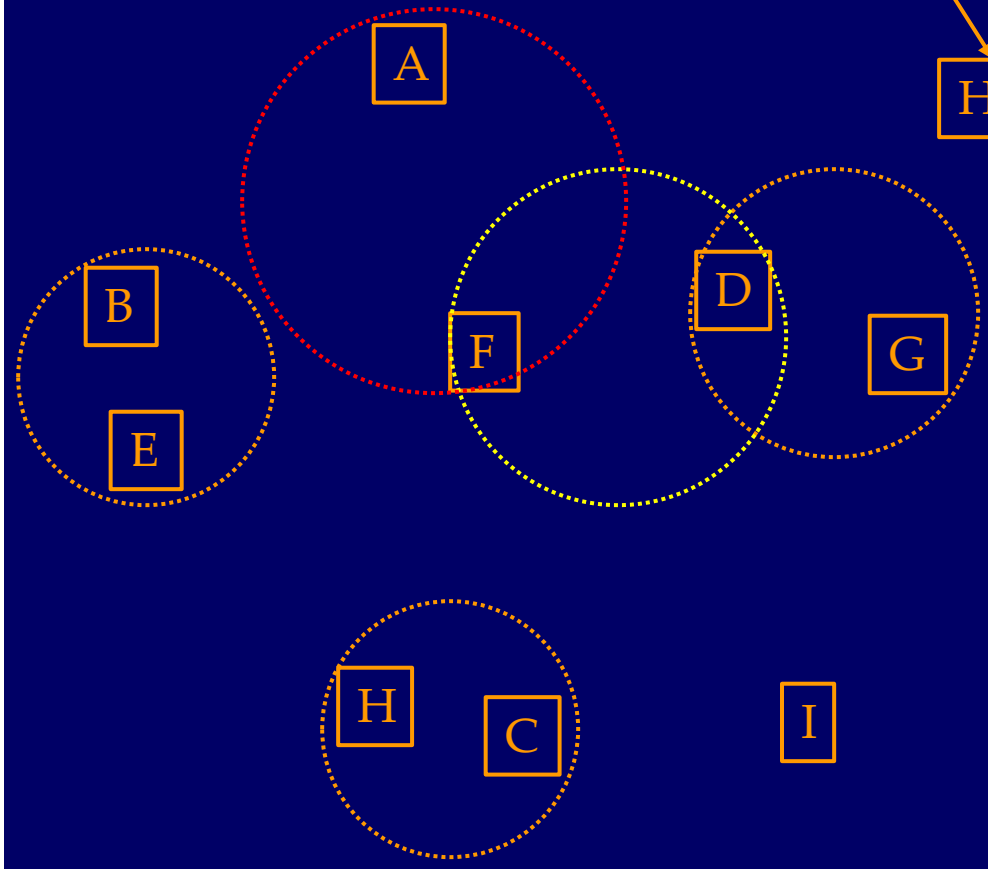
single linkage

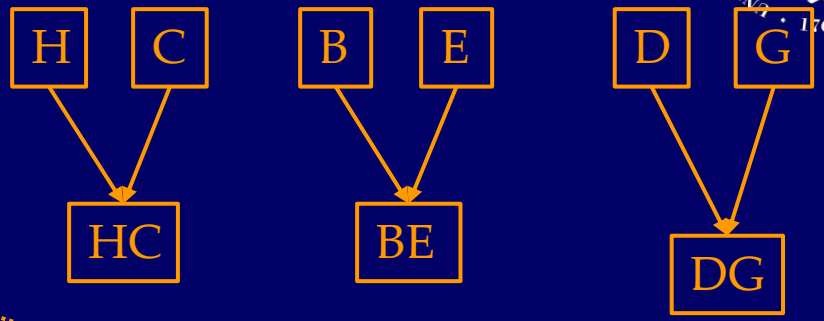




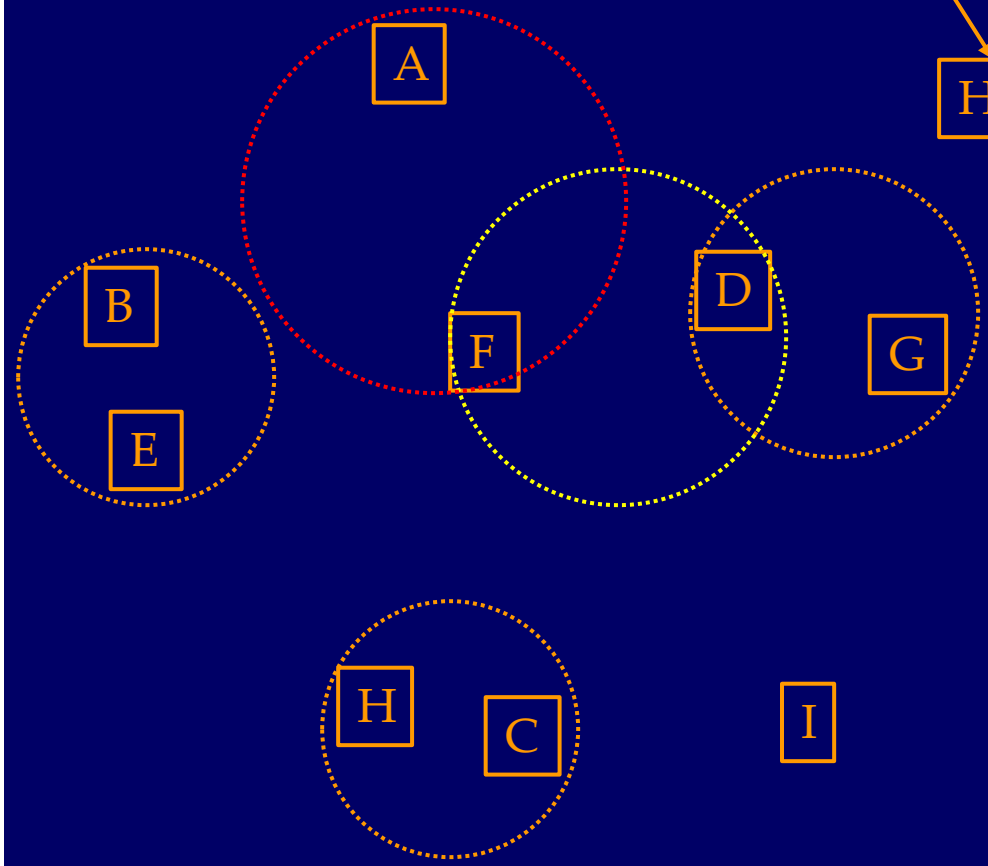

single linkage
A
B
E
F
D
G
H
C
I
I C H E B A F D G
CH
EB
DG
ICH
FDG
AFDG
EBAFDG
ICHEBAFDG

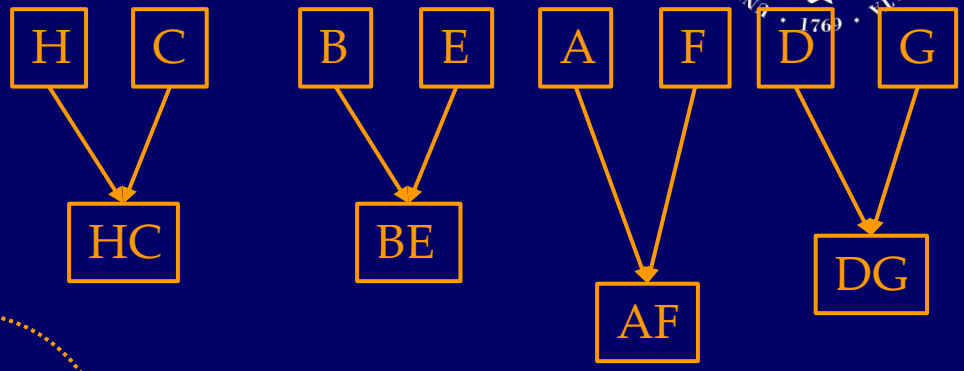
complete linkage



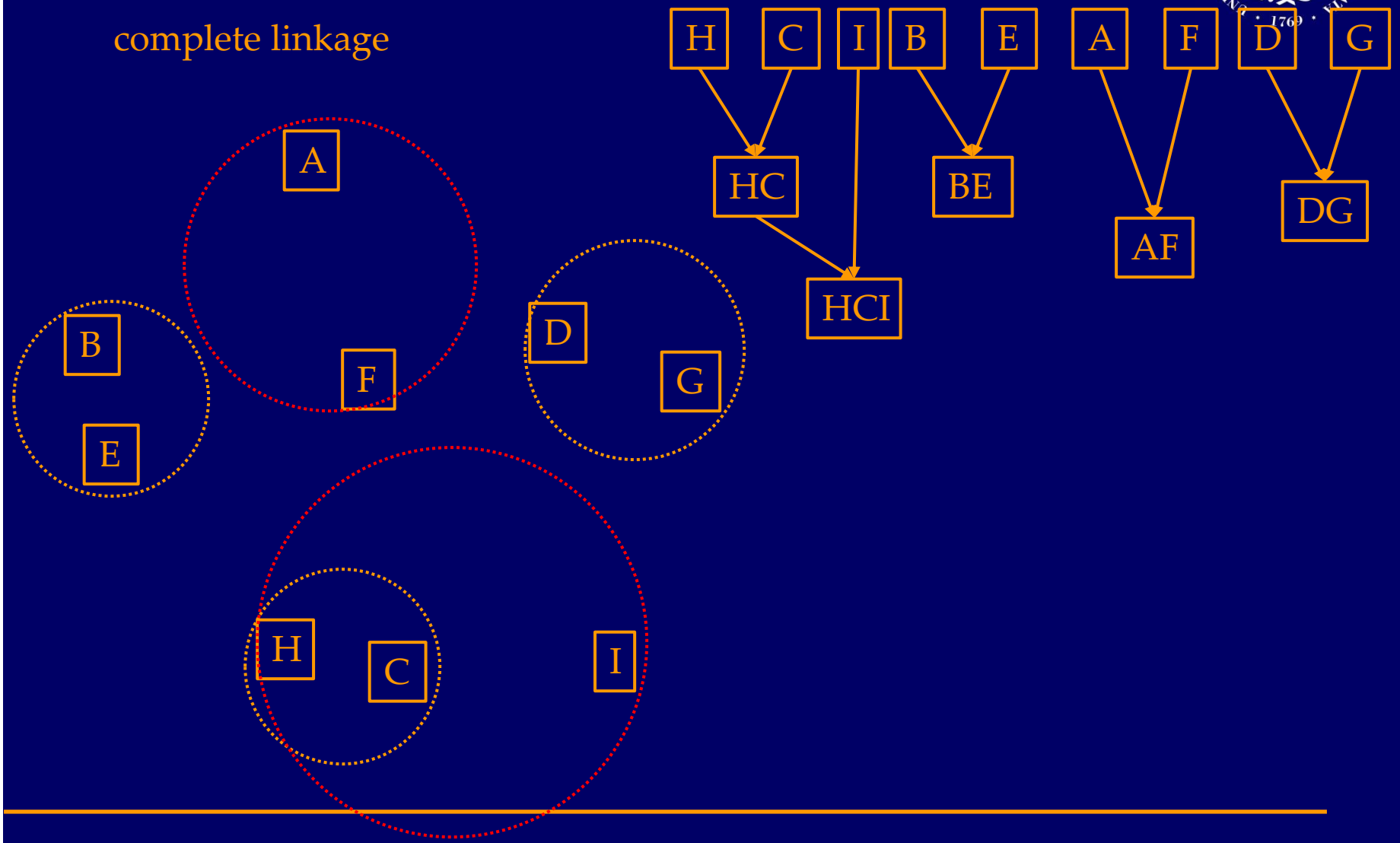


complete linkage

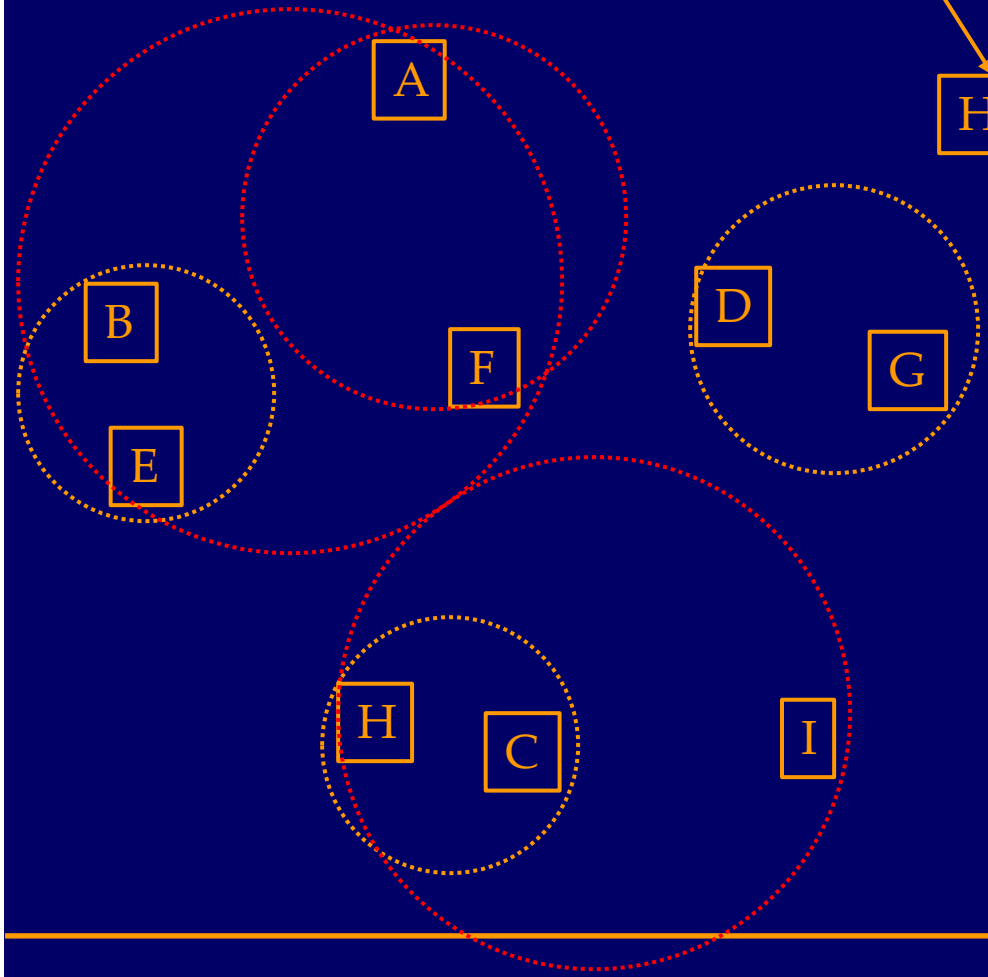


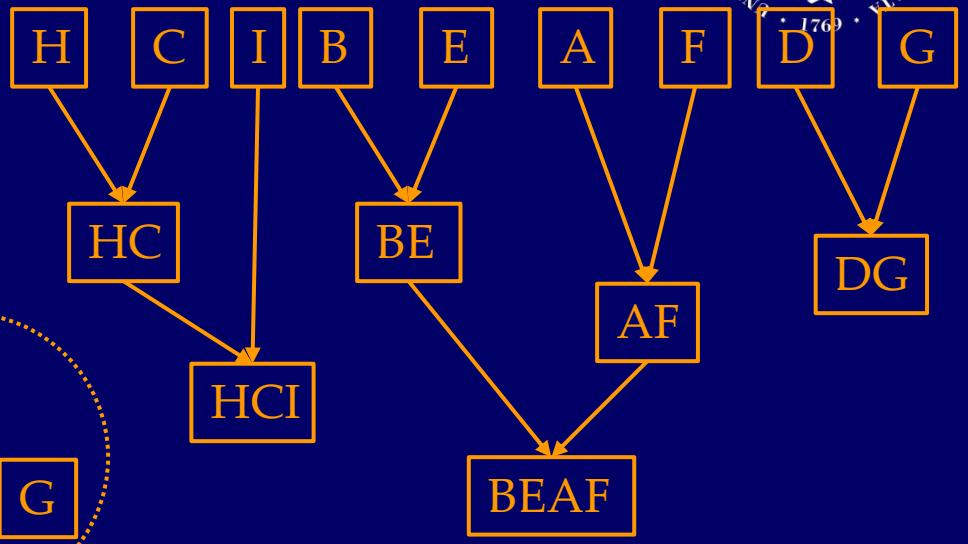


complete linkage

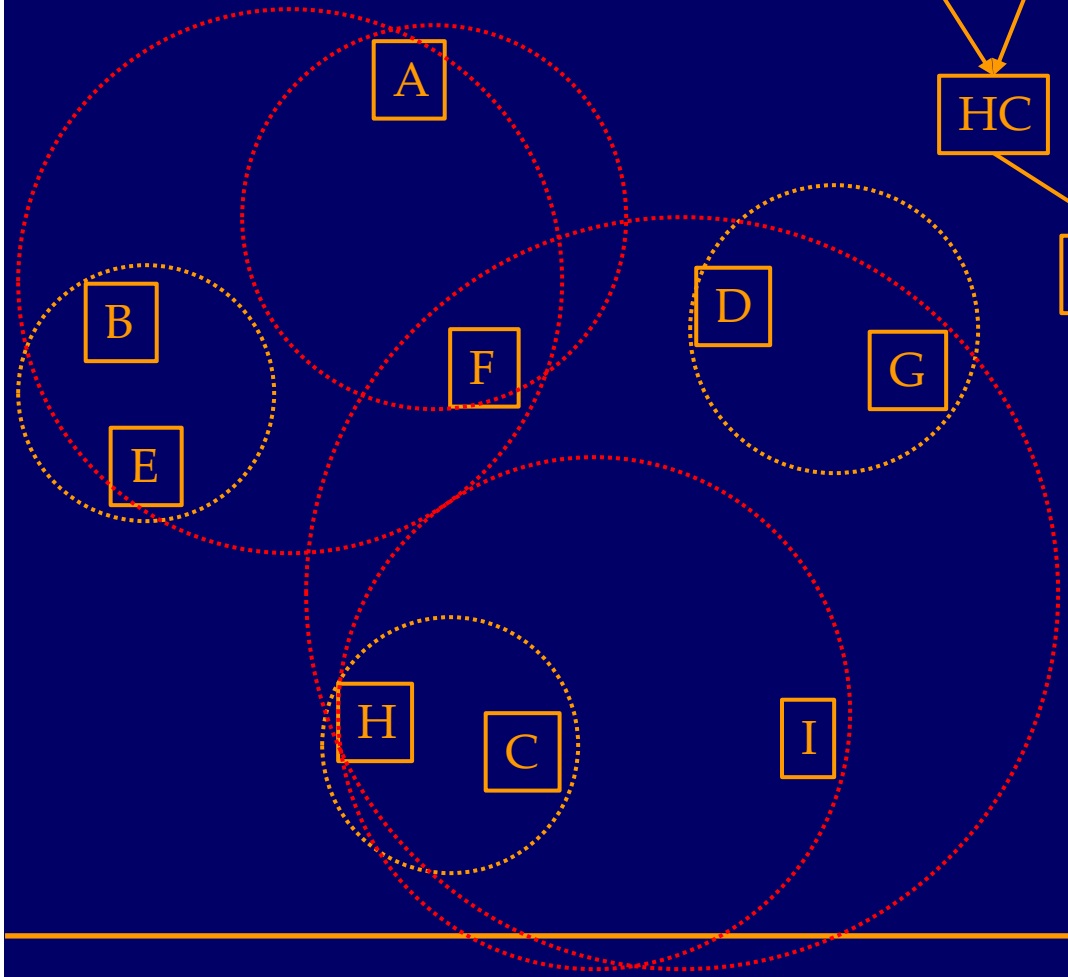


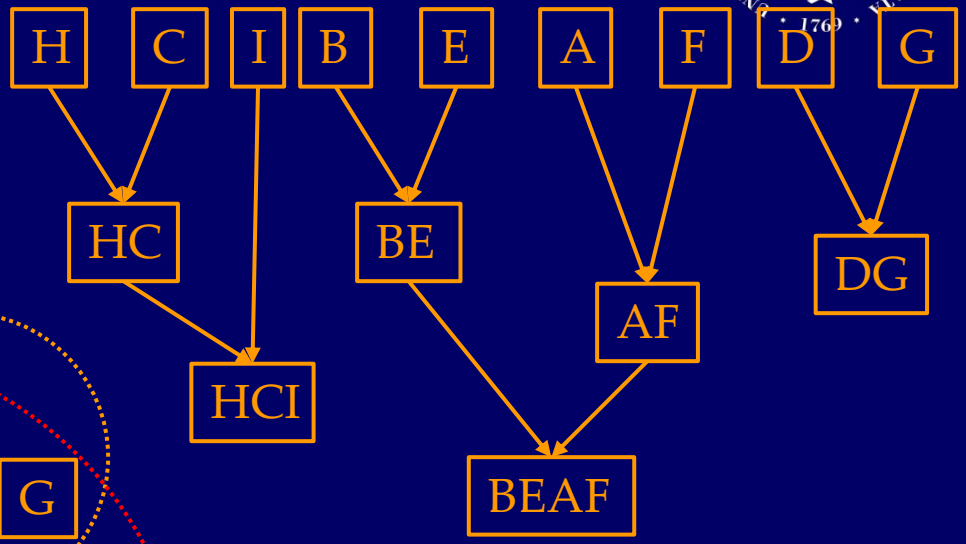
complete linkage



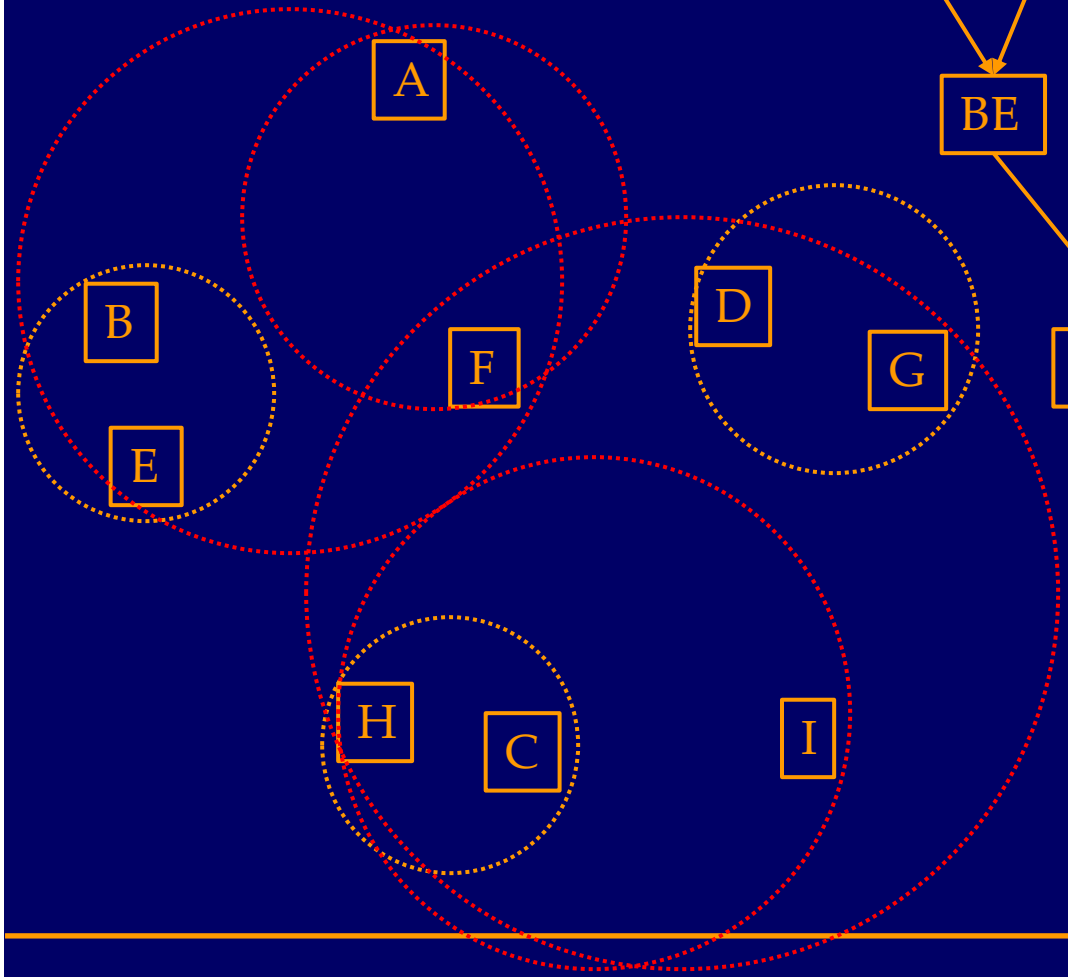


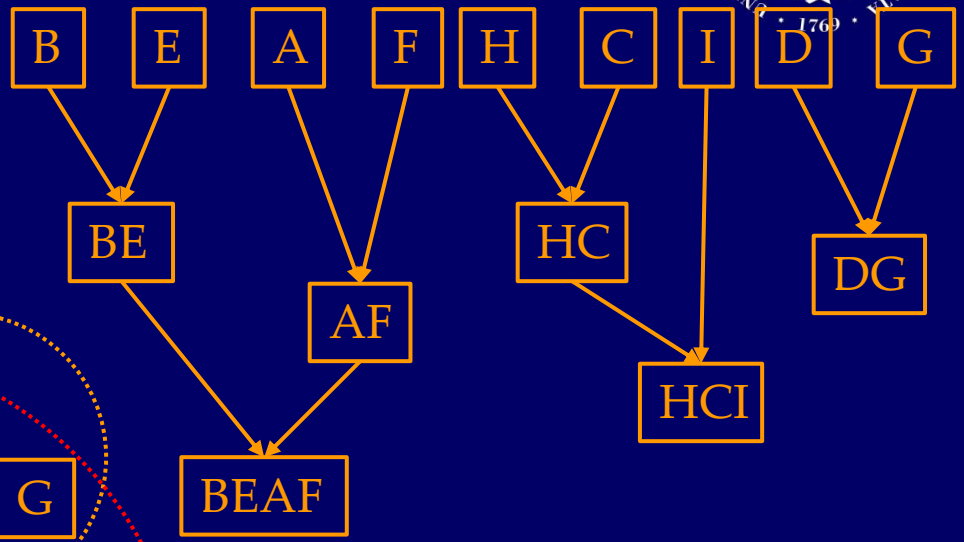
complete linkage



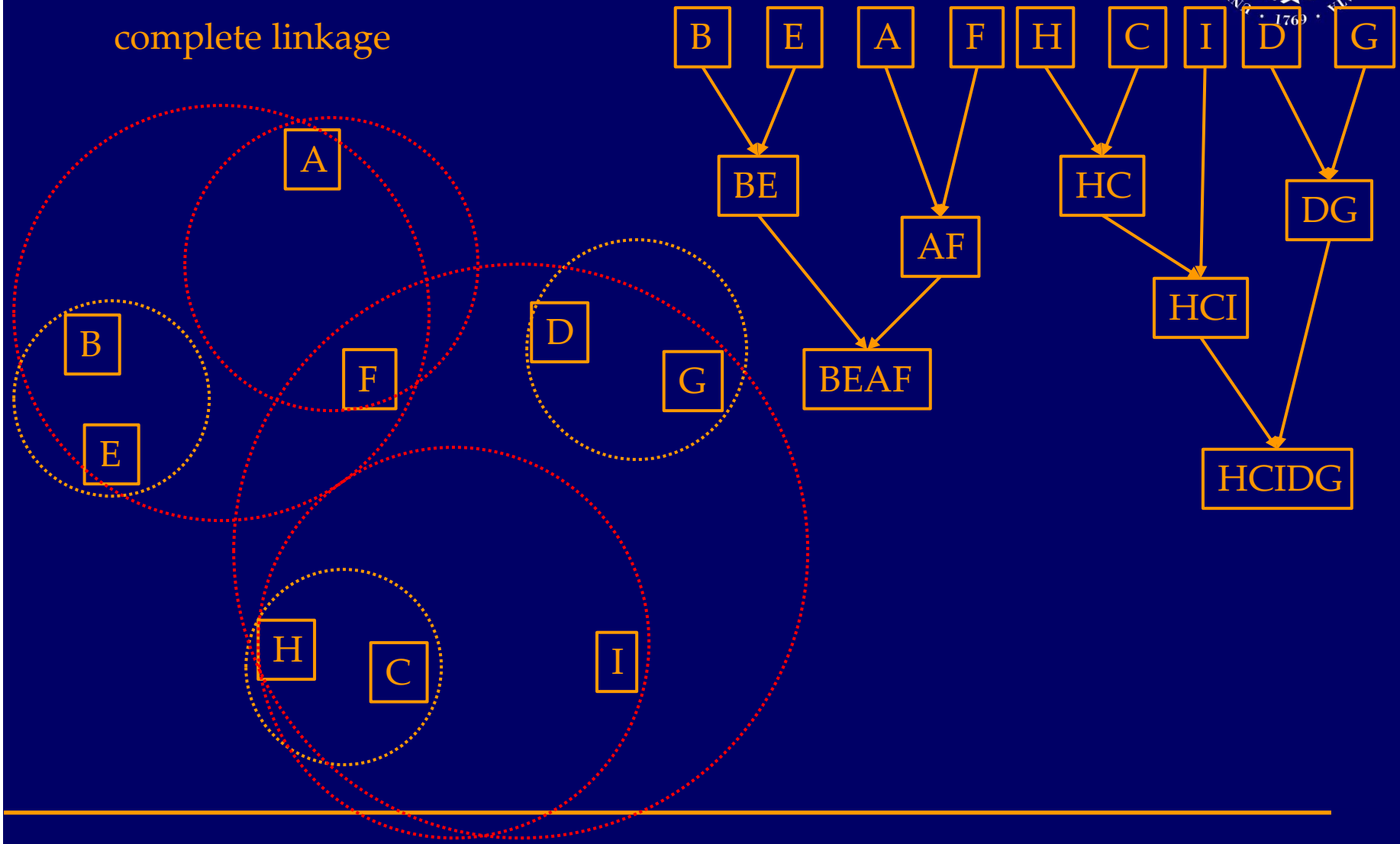


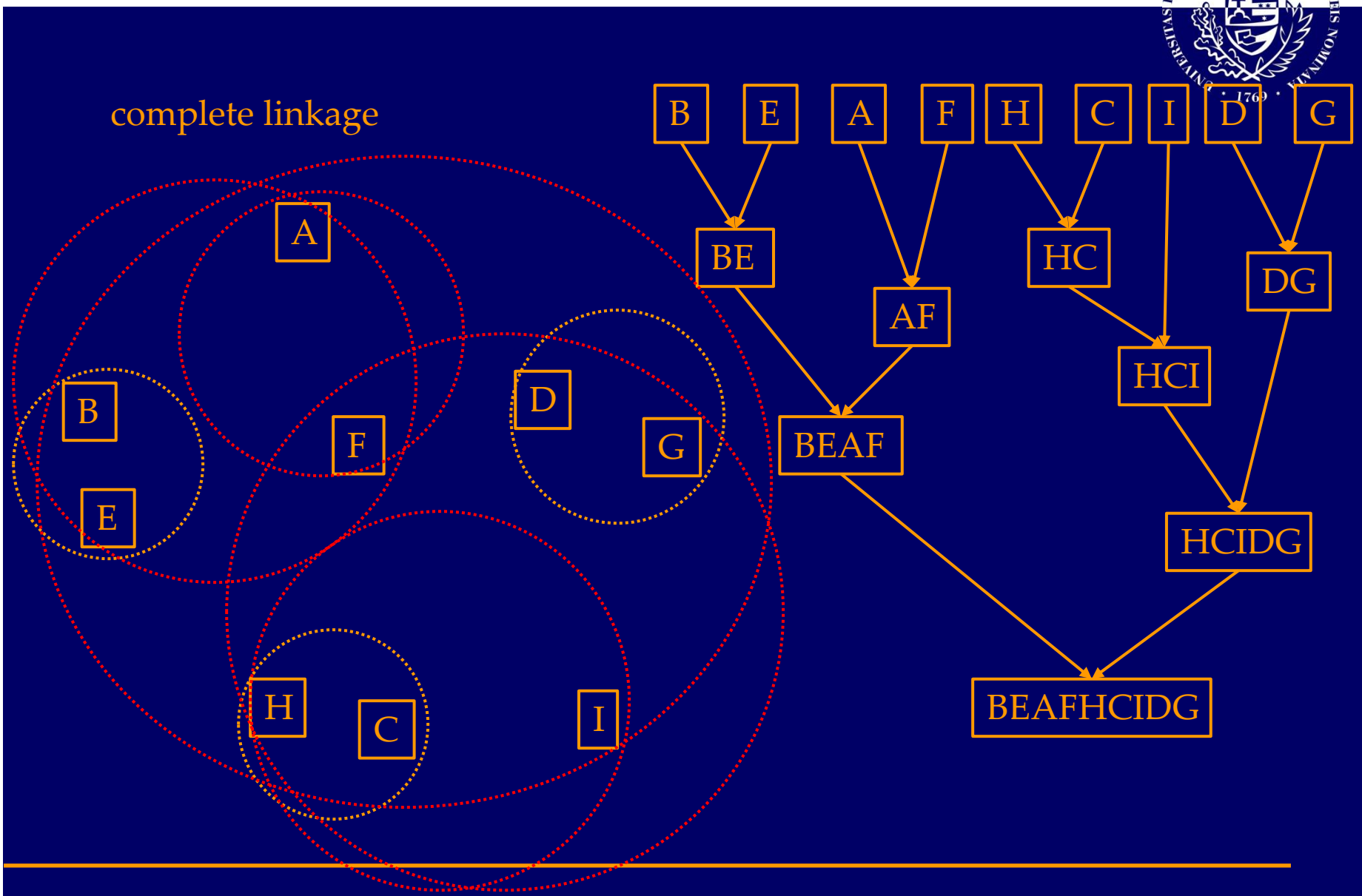
complete linkage



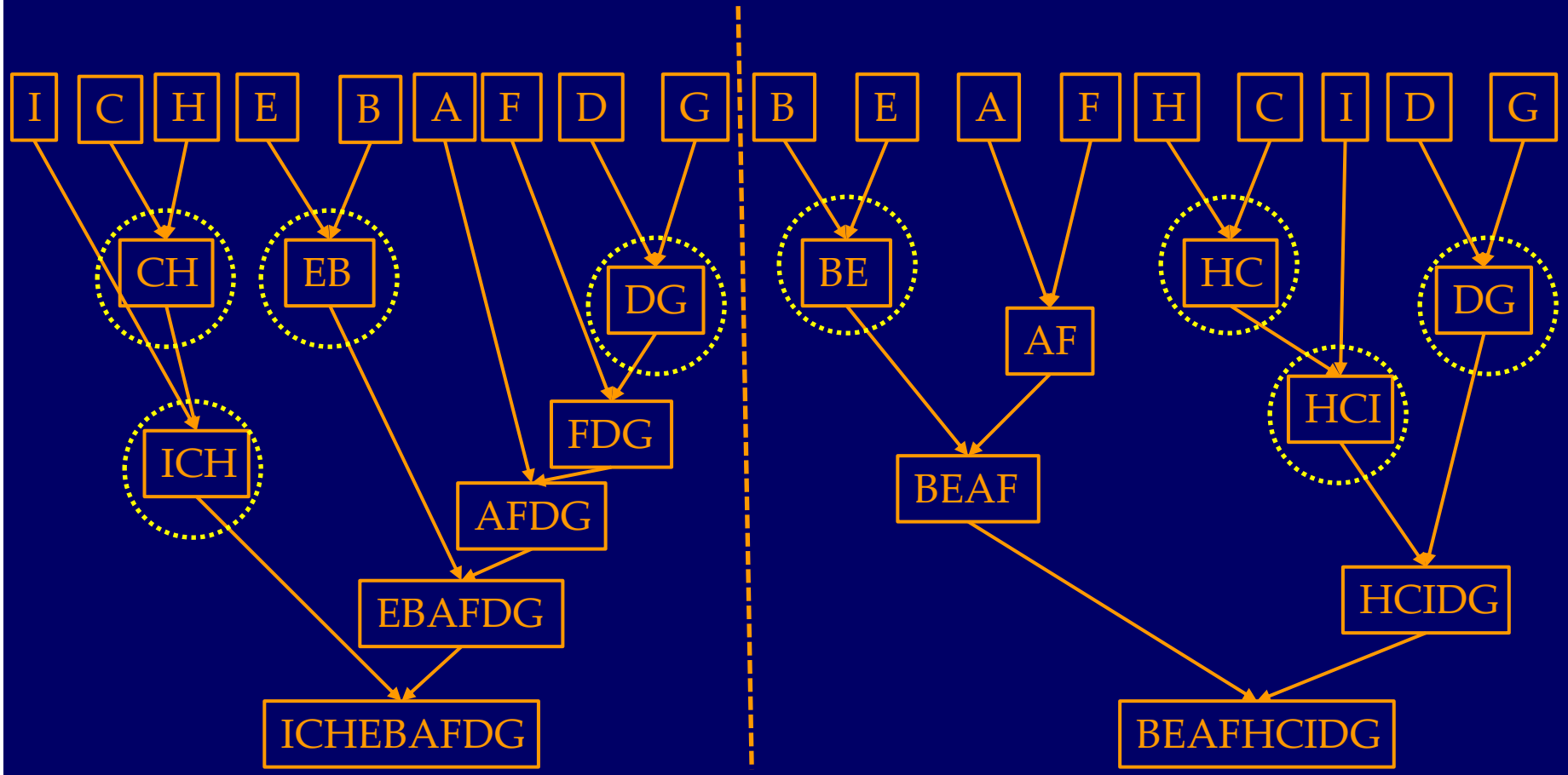


complete linkage



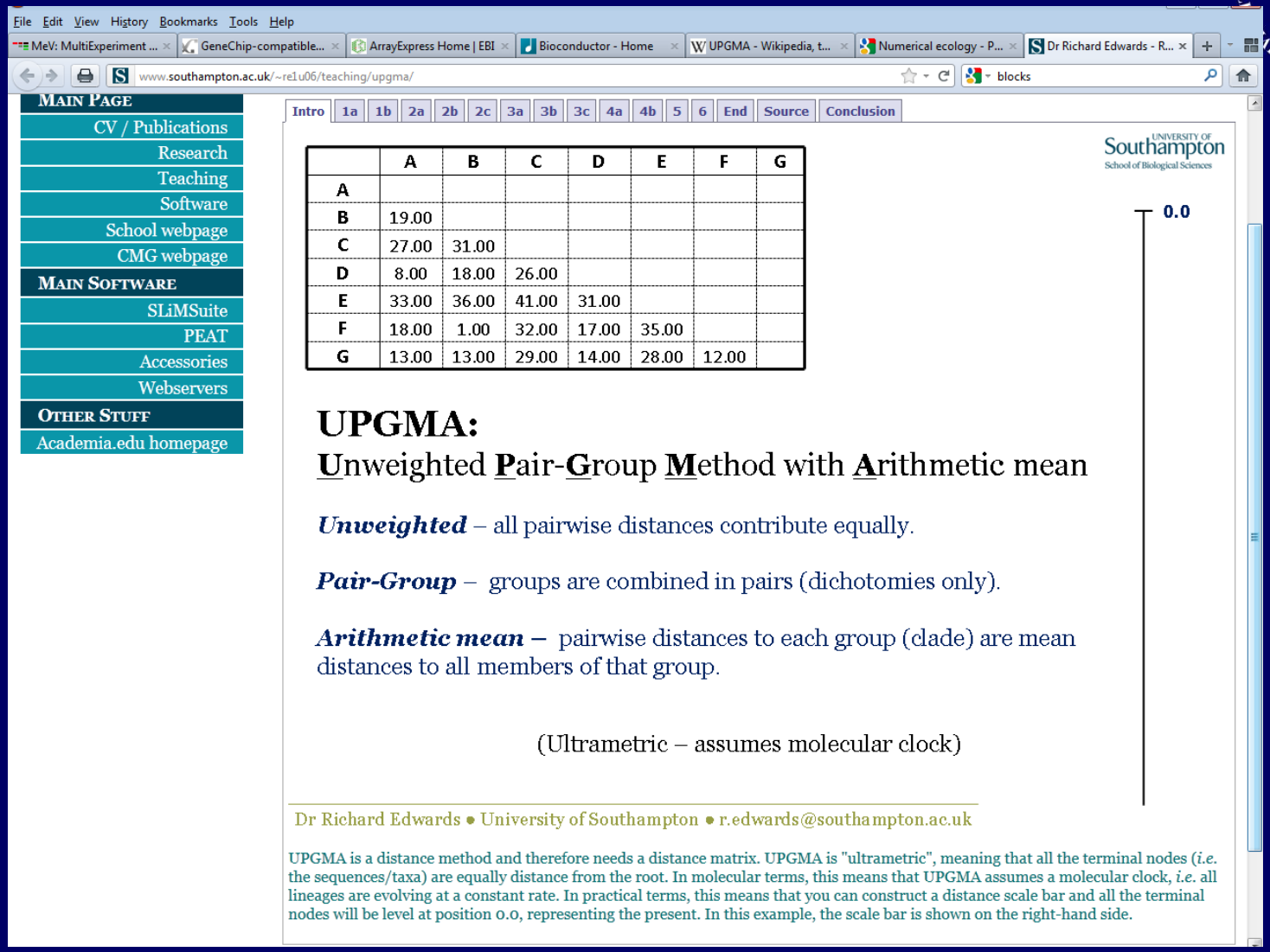

complete linkage
A
B
E
F
D
G
H
C
I
B
E
A
F
H
C
I
D
G
BE
AF
HC
DG
HCI
BEAF
HCIDG
BEAFHCIDG

Single vs. complete

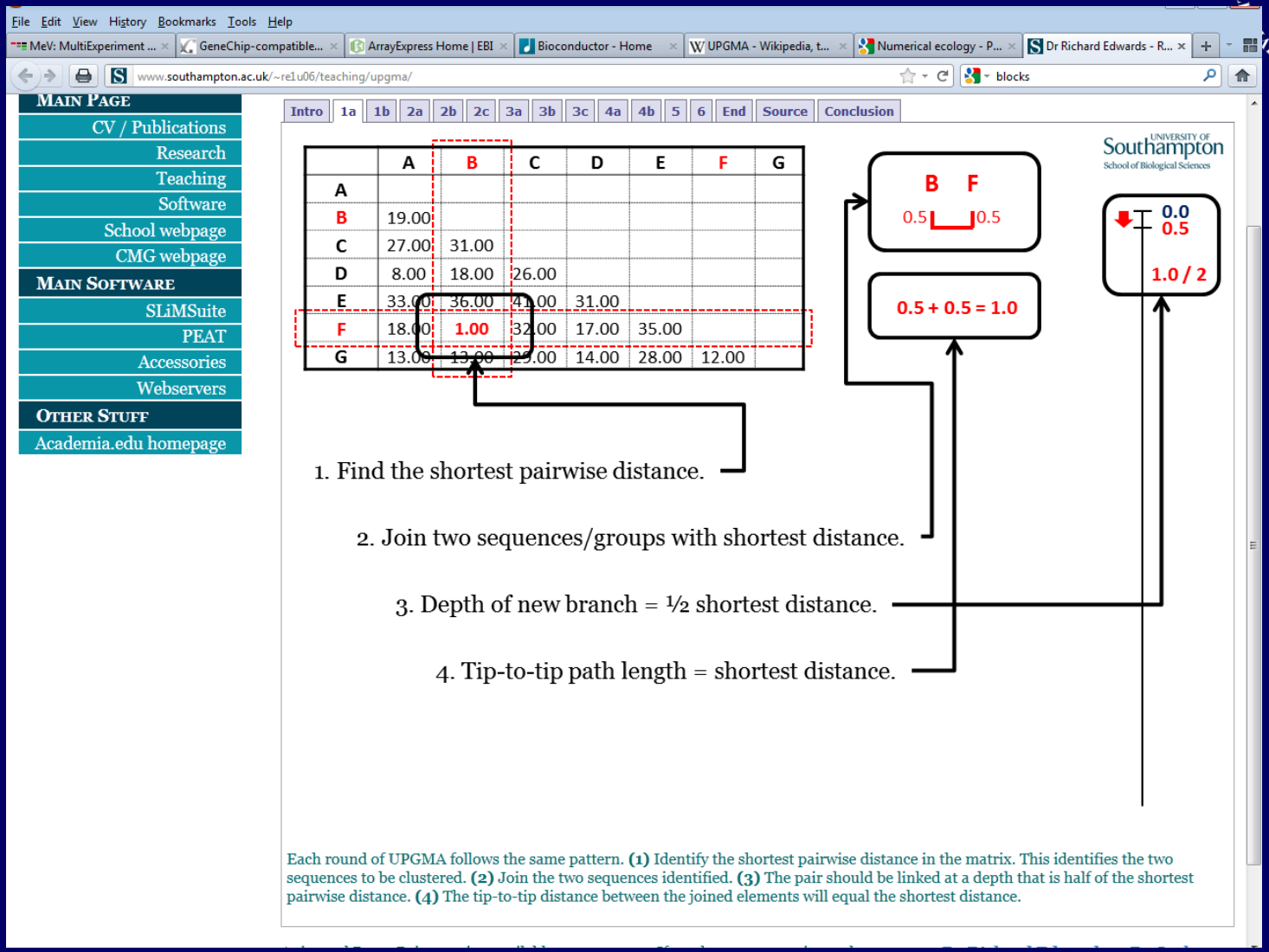


UPGMA

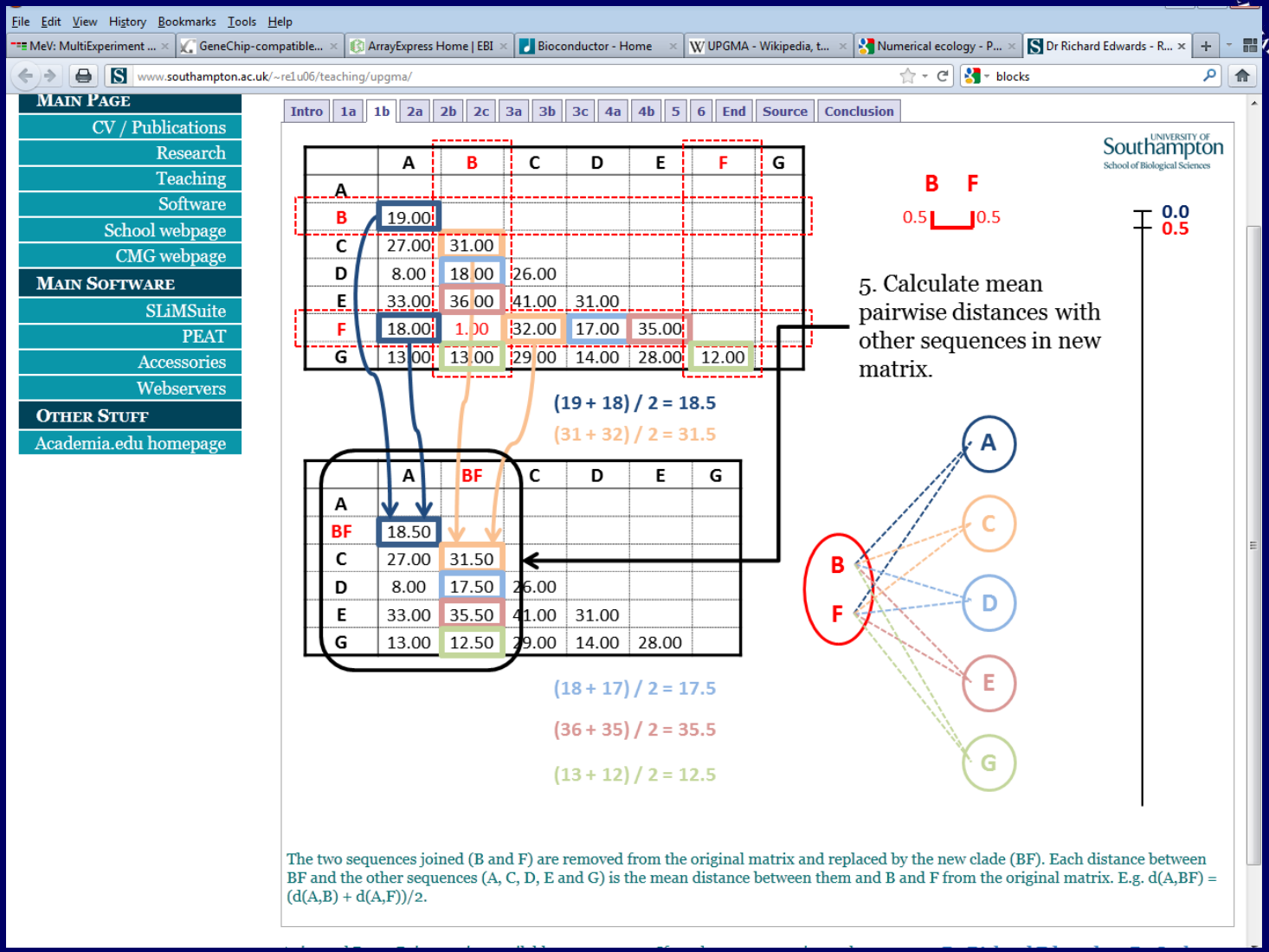
- "Unweighted Pair-Group Method with Aritmetic mean"
- Általános módszer, bármilyen távolságmatrixon működik
- Példa:
  - http://www.slimsuite.unsw.edu.au/teaching/upgma/

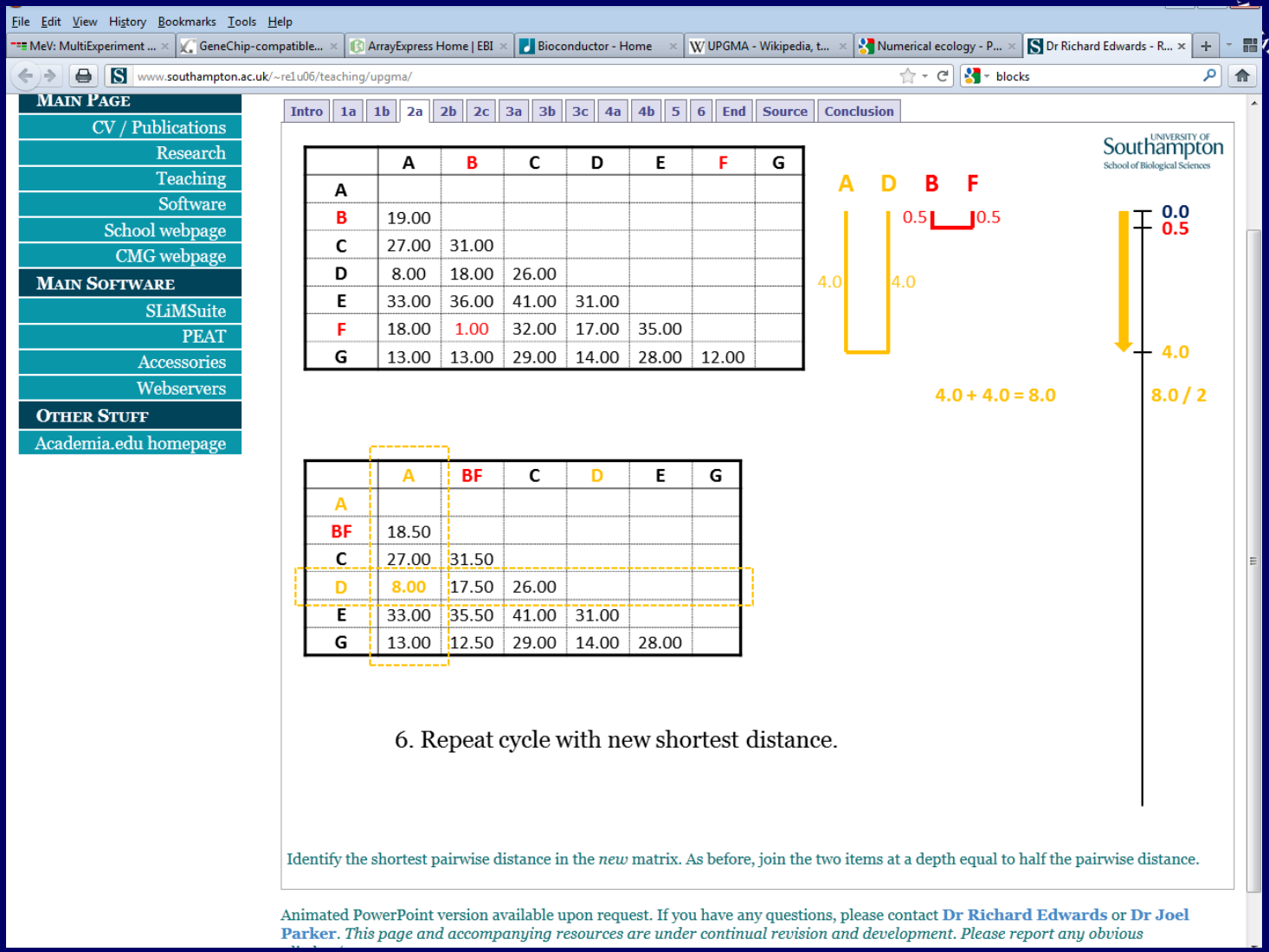


	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	



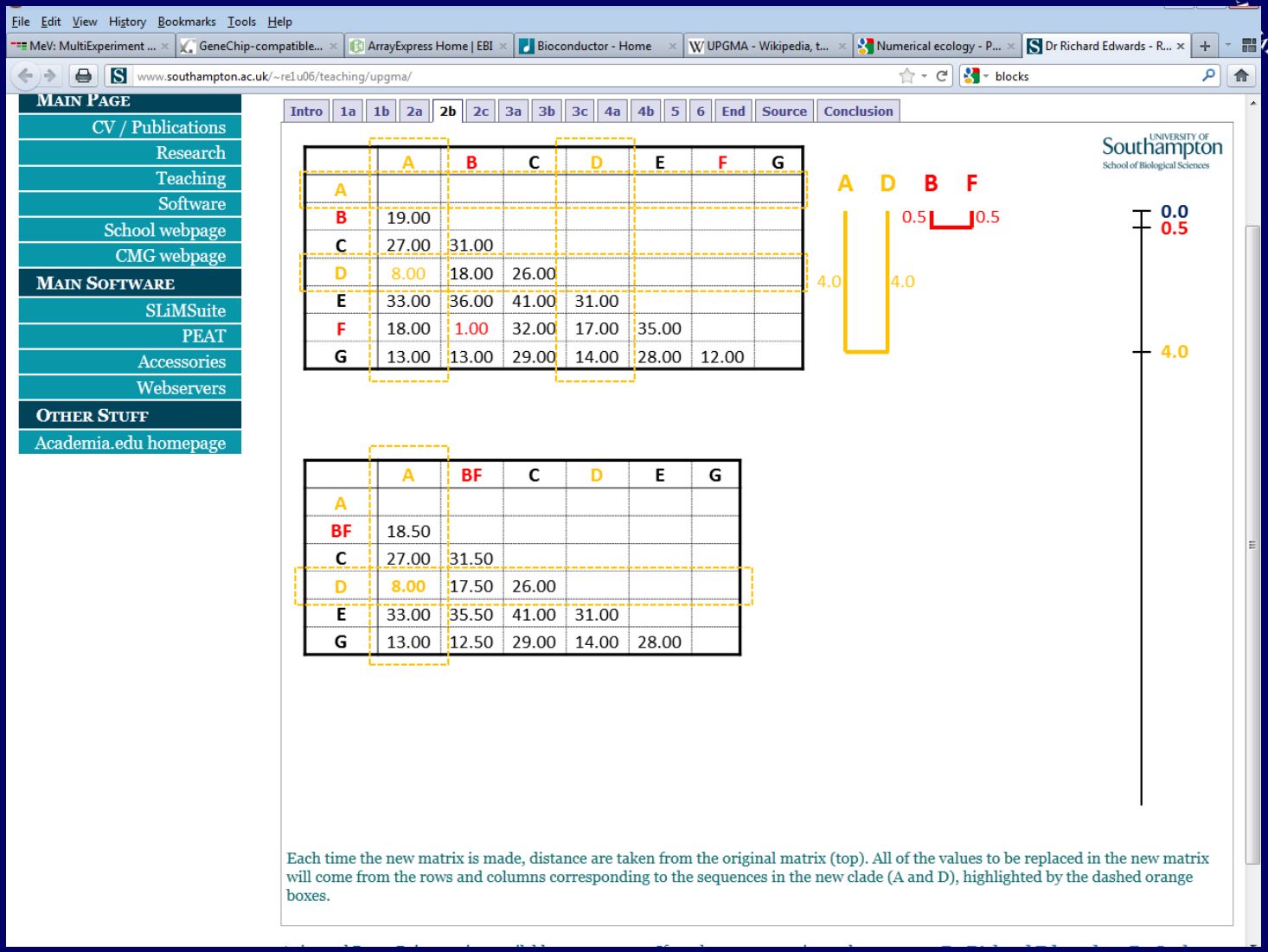
	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	


File Edit View History Bookmarks Tools Help
www.southampton.ac.uk/~re1u06/teaching/upgma/
MAIN PAGE
CV / Publications
Research
Teaching
Software
School webpage
CMG webpage
MAIN SOFTWARE
SLiMSuite
PEAT
Accessories
Webservers
OTHER STUFF
Academia.edu homepage
Intro 1a 1b 2a 2b 2c 3a 3b 3c 4a 4b 5 6 End Source Conclusion
University of Southampton
School of Biological Sciences
5. Calculate mean pairwise distances with other sequences in new matrix.
(19 + 18) / 2 = 18.5
(31 + 32) / 2 = 31.5
(18 + 17) / 2 = 17.5
(36 + 35) / 2 = 35.5
(13 + 12) / 2 = 12.5
0.0
0.5
The two sequences joined (B and F) are removed from the original matrix and replaced by the new clade (BF). Each distance between BF and the other sequences (A, C, D, E and G) is the mean distance between them and B and F from the original matrix. E.g. d(A,BF) = (d(A,B) + d(A,F))/2.



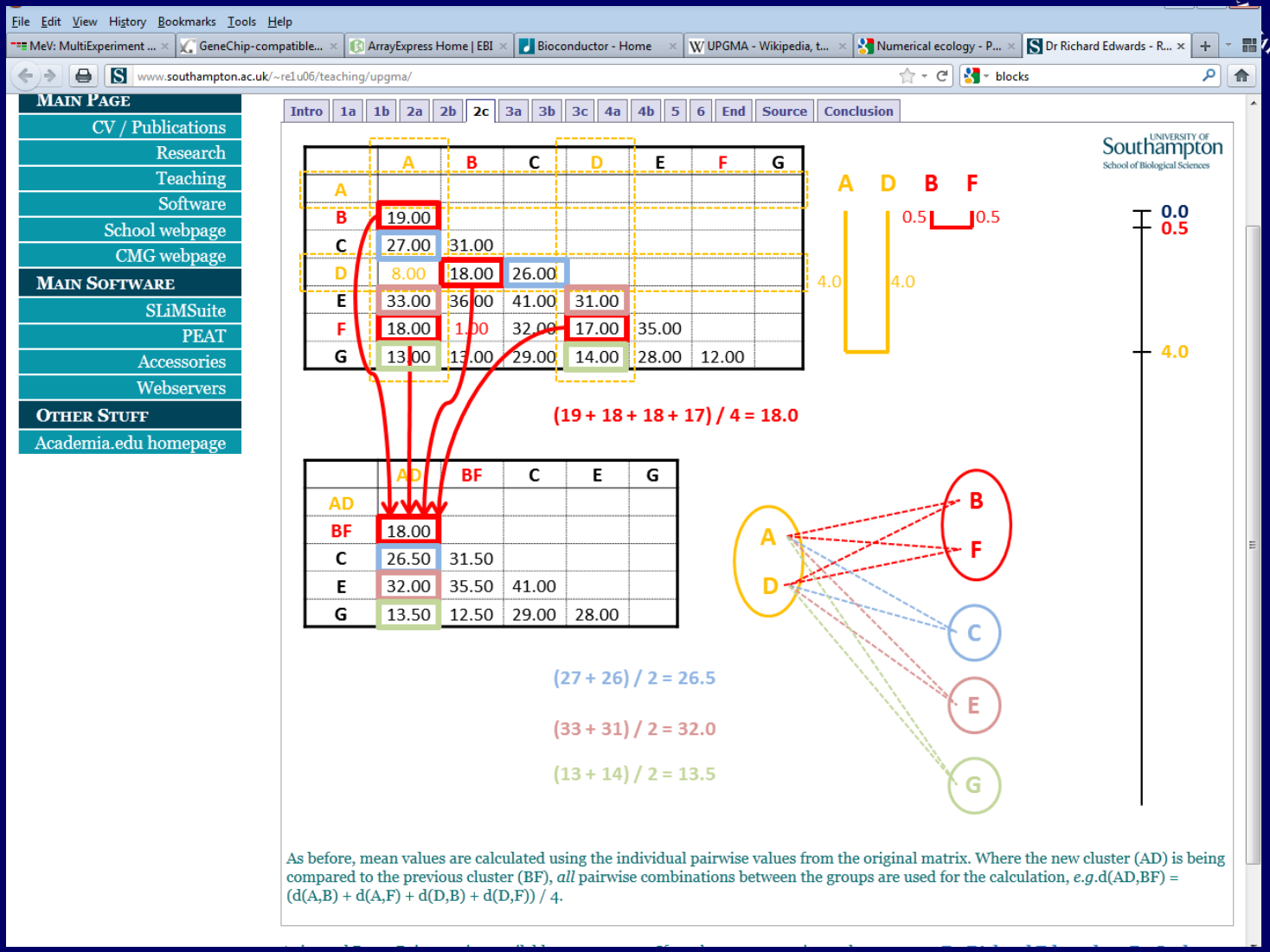
	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

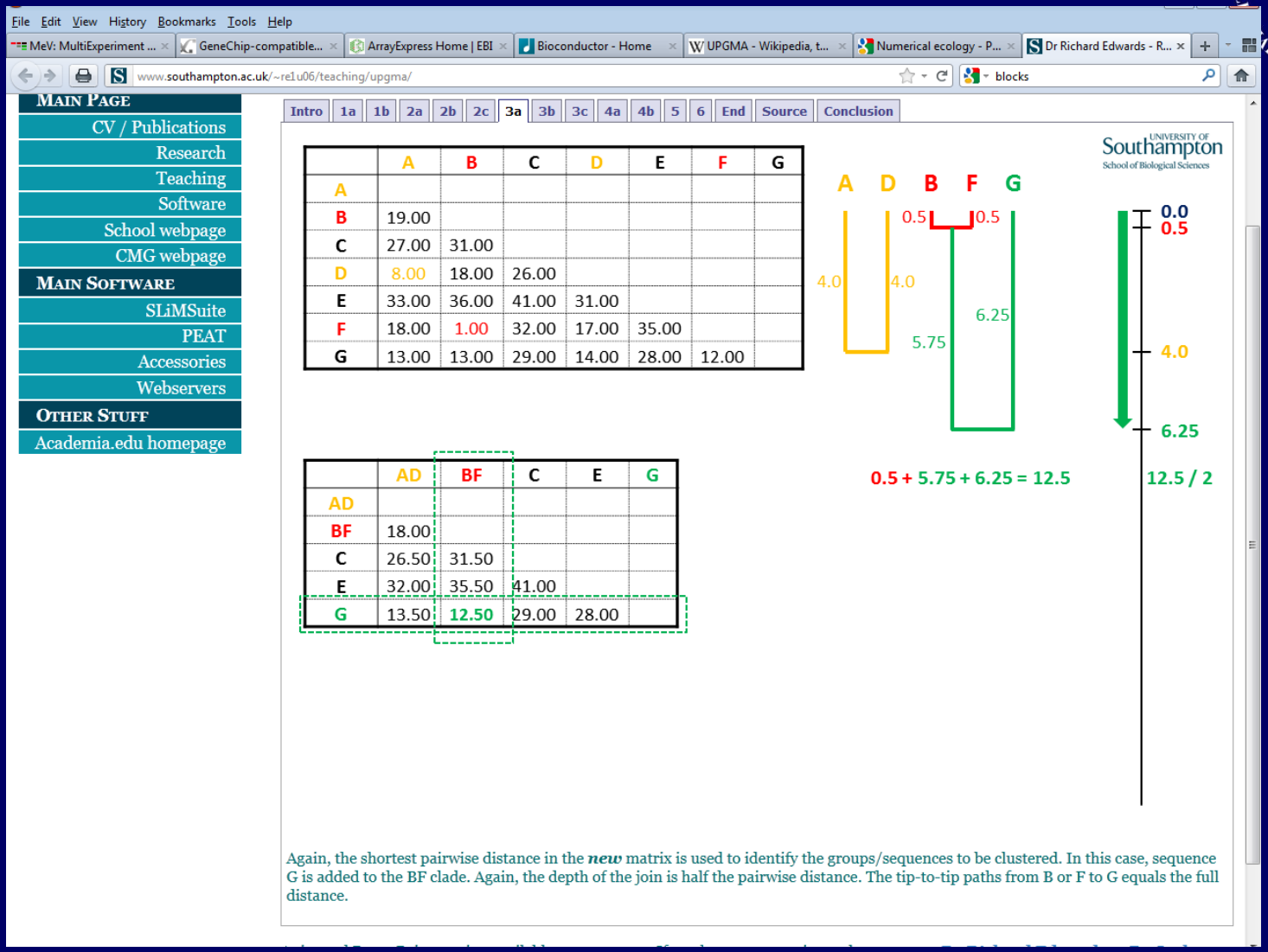
	A	BF	C	D	E	G
A						
BF	18.50					
C	27.00	31.50				
D	8.00	17.50	26.00			
E	33.00	35.50	41.00	31.00		
G	13.00	12.50	29.00	14.00	28.00	



	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

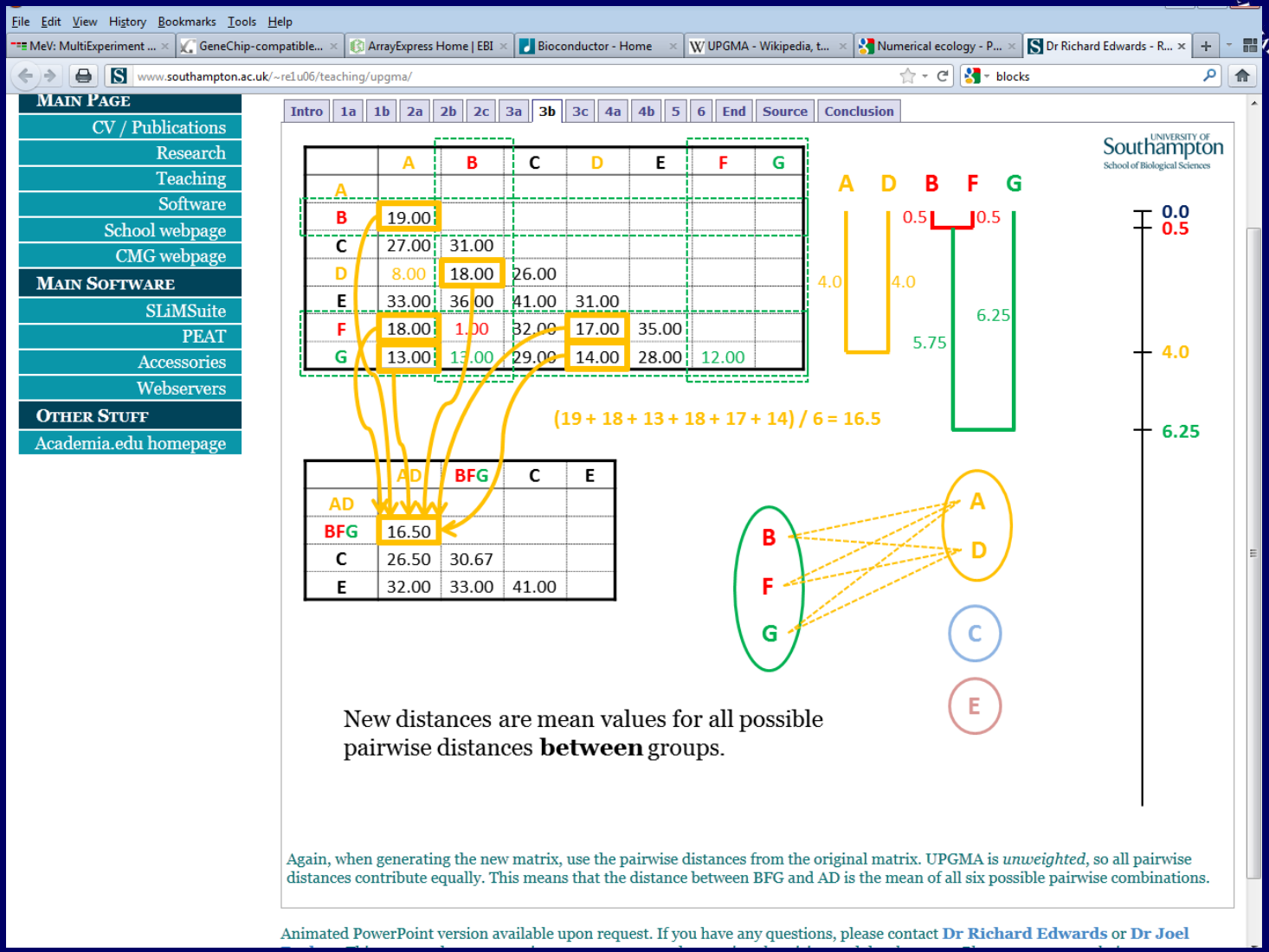
	A	BF	C	D	E	G
A						
BF	18.50					
C	27.00	31.50				
D	8.00	17.50	26.00			
E	33.00	35.50	41.00	31.00		
G	13.00	12.50	29.00	14.00	28.00	

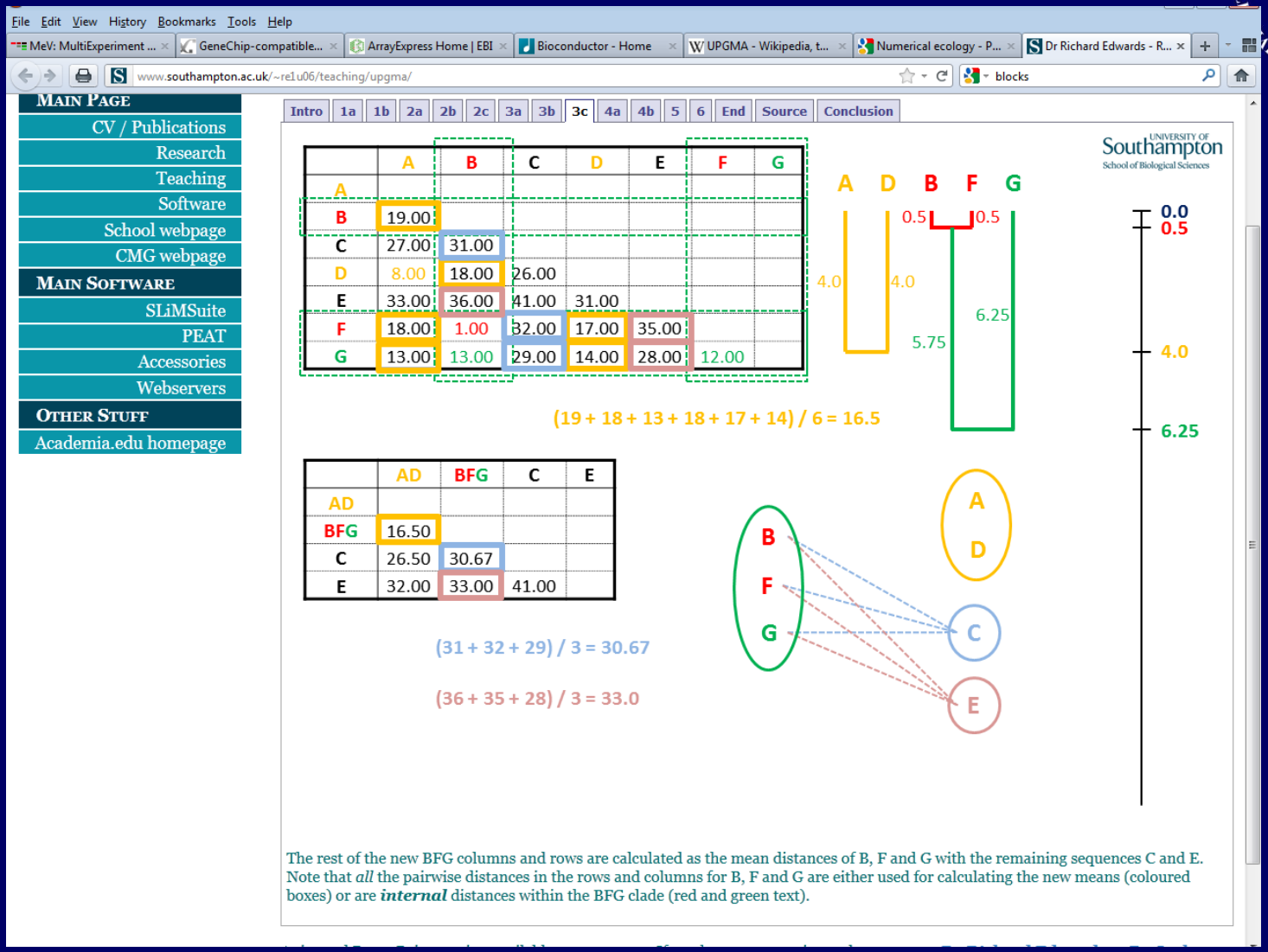

File Edit View History Bookmarks Tools Help
www.southampton.ac.uk/~re1u06/teaching/upgma/
MAIN PAGE
CV / Publications
Research
Teaching
Software
School webpage
CMG webpage
MAIN SOFTWARE
SLiMSuite
PEAT
Accessories
Webservers
OTHER STUFF
Academia.edu homepage
Intro 1a 1b 2a 2b 2c 3a 3b 3c 4a 4b 5 6 End Source Conclusion
University of Southampton
School of Biological Sciences
(19 + 18 + 18 + 17) / 4 = 18.0
(27 + 26) / 2 = 26.5
(33 + 31) / 2 = 32.0
(13 + 14) / 2 = 13.5
As before, mean values are calculated using the individual pairwise values from the original matrix. Where the new cluster (AD) is being compared to the previous cluster (BF), all pairwise combinations between the groups are used for the calculation, e.g.d(AD,BF) = (d(A,B) + d(A,F) + d(D,B) + d(D,F)) / 4.



	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

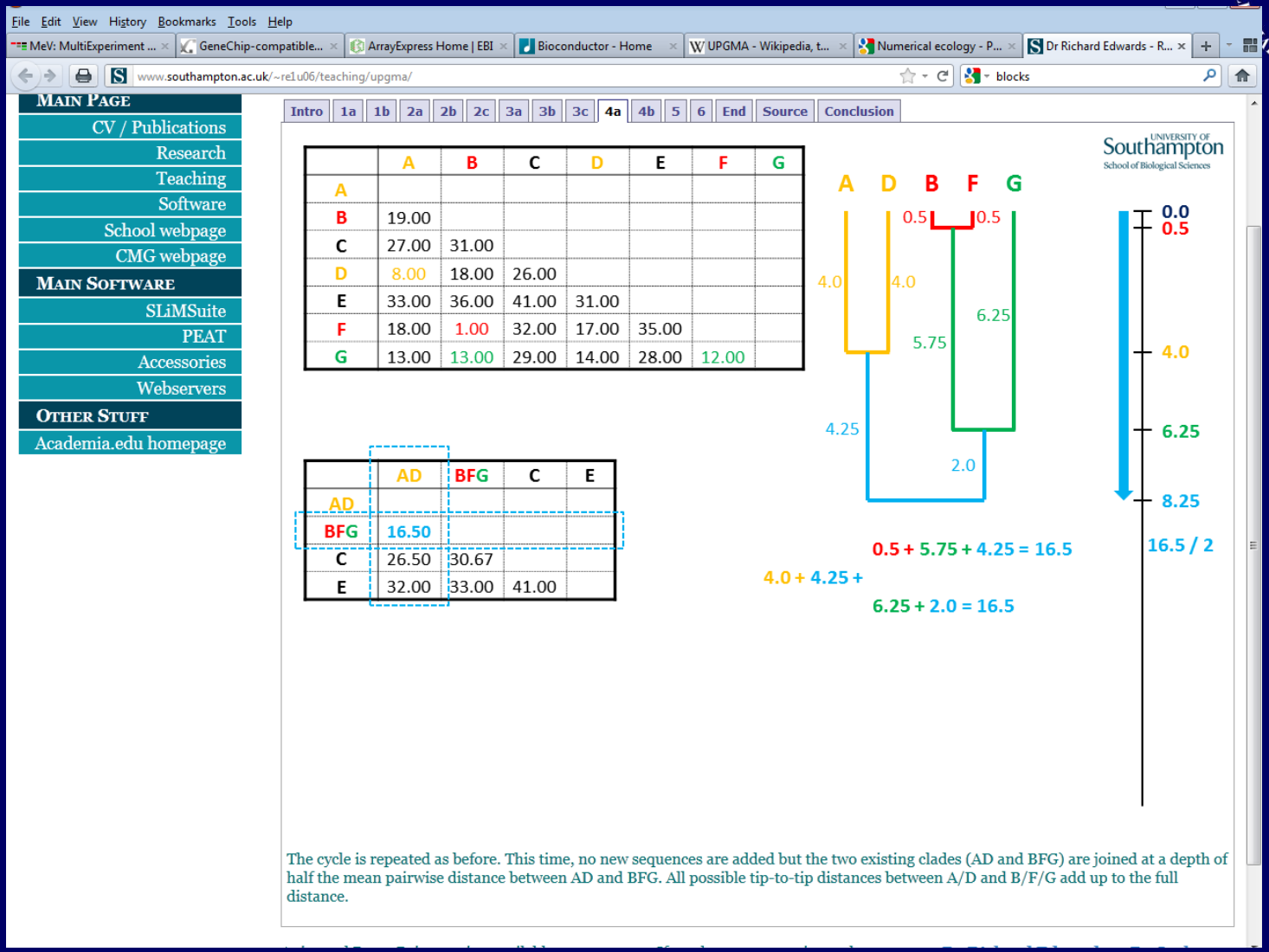
	AD	BF	C	E	G
AD					
BF	18.00				
C	26.50	31.50			
E	32.00	35.50	41.00		
G	13.50	12.50	29.00	28.00	


File Edit View History Bookmarks Tools Help
www.southampton.ac.uk/~re1u06/teaching/upgma/
MAIN PAGE
CV / Publications
Research
Teaching
Software
School webpage
CMG webpage
MAIN SOFTWARE
SLiMSuite
PEAT
Accessories
Webservers
OTHER STUFF
Academia.edu homepage
Intro 1a 1b 2a 2b 2c 3a 3b 3c 4a 4b 5 6 End Source Conclusion
University of Southampton
School of Biological Sciences
(19 + 18 + 13 + 18 + 17 + 14) / 6 = 16.5
New distances are mean values for all possible pairwise distances between groups.
Again, when generating the new matrix, use the pairwise distances from the original matrix. UPGMA is unweighted, so all pairwise distances contribute equally. This means that the distance between BFG and AD is the mean of all six possible pairwise combinations.
Animated PowerPoint version available upon request. If you have any questions, please contact Dr Richard Edwards or Dr Joel



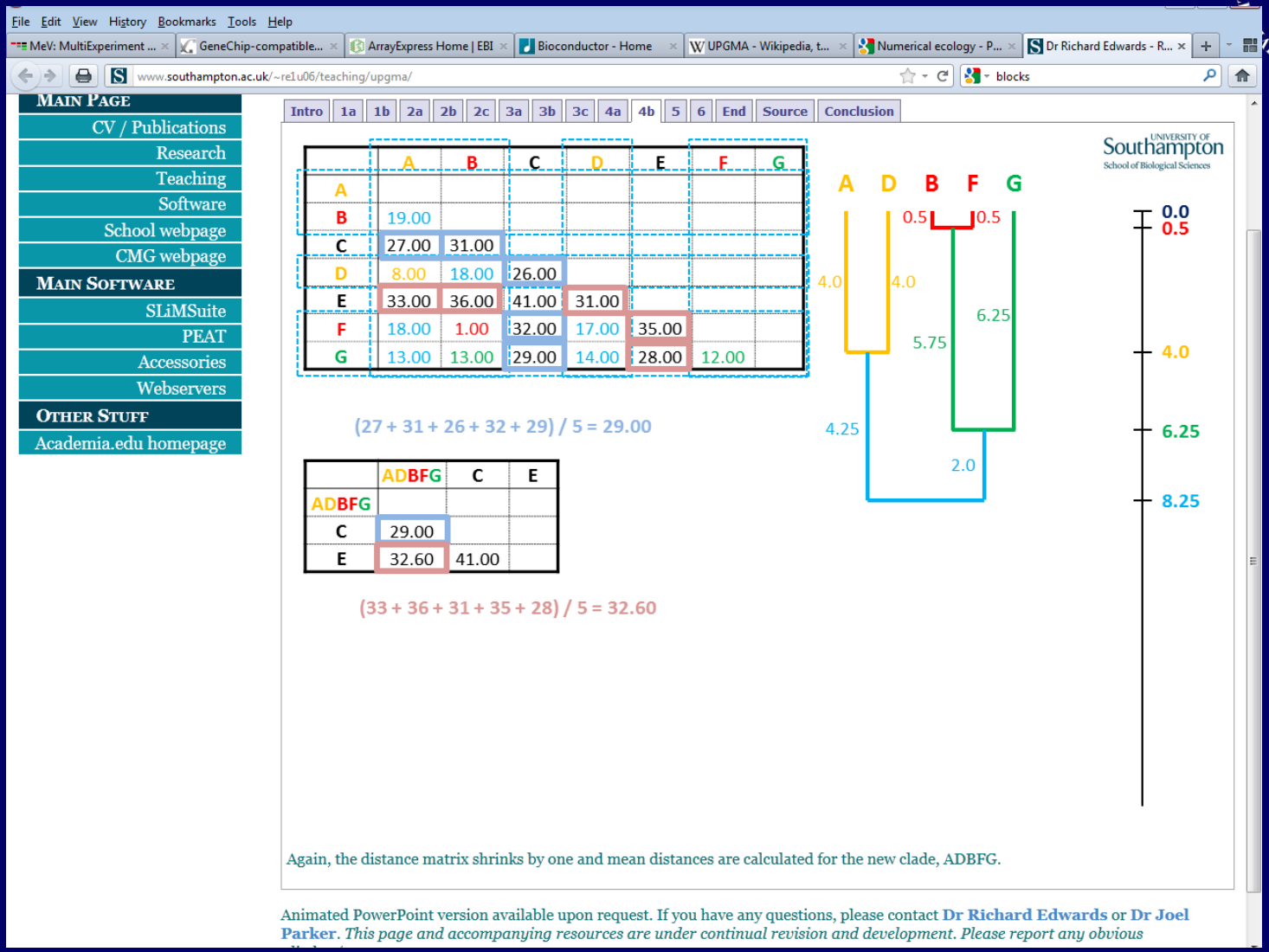
	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

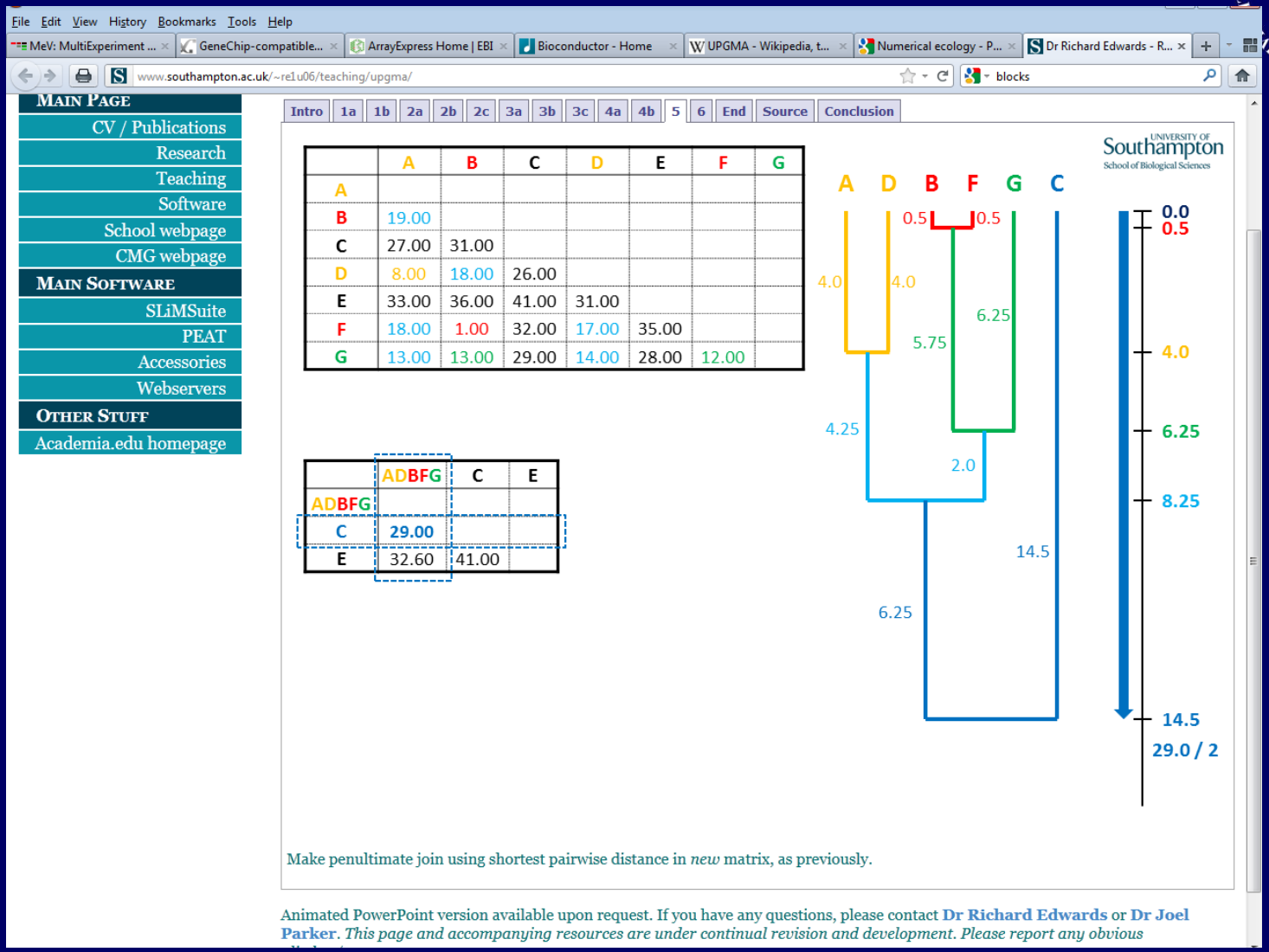
	AD	BFG	C	E
AD				
BFG	16.50			
C	26.50	30.67		
E	32.00	33.00	41.00	



	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

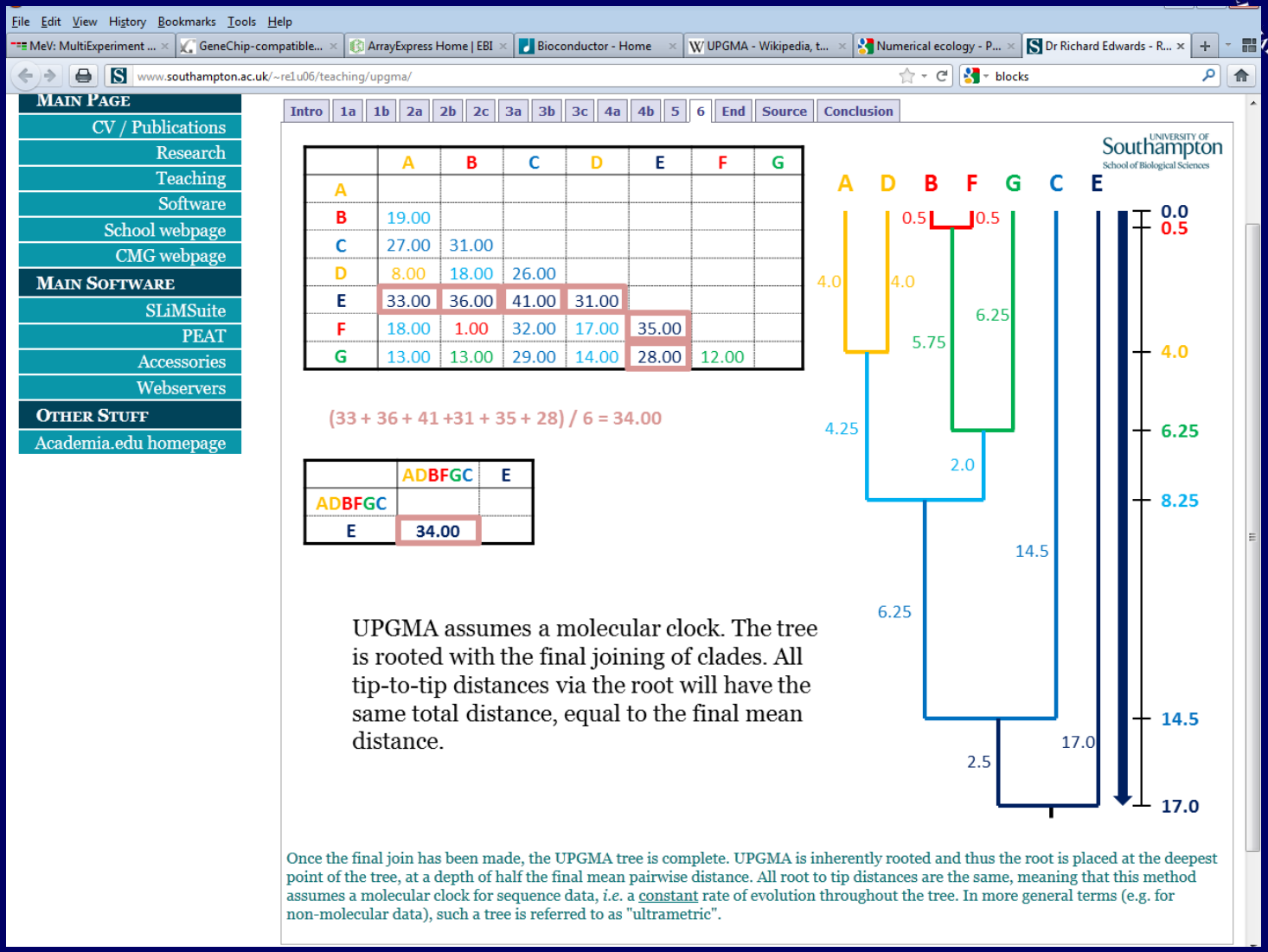
	AD	BFG	C	E
AD				
BFG	16.50			
C	26.50	30.67		
E	32.00	33.00	41.00	


File Edit View History Bookmarks Tools Help
www.southampton.ac.uk/~re1u06/teaching/upgma/
MAIN PAGE
CV / Publications
Research
Teaching
Software
School webpage
CMG webpage
MAIN SOFTWARE
SLiMSuite
PEAT
Accessories
Webservers
OTHER STUFF
Academia.edu homepage
Intro 1a 1b 2a 2b 2c 3a 3b 3c 4a 4b 5 6 End Source Conclusion
University of Southampton
School of Biological Sciences
(27 + 31 + 26 + 32 + 29) / 5 = 29.00
(33 + 36 + 31 + 35 + 28) / 5 = 32.60
Again, the distance matrix shrinks by one and mean distances are calculated for the new clade, ADBFG.
Animated PowerPoint version available upon request. If you have any questions, please contact Dr Richard Edwards or Dr Joel Parker. This page and accompanying resources are under continual revision and development. Please report any obvious



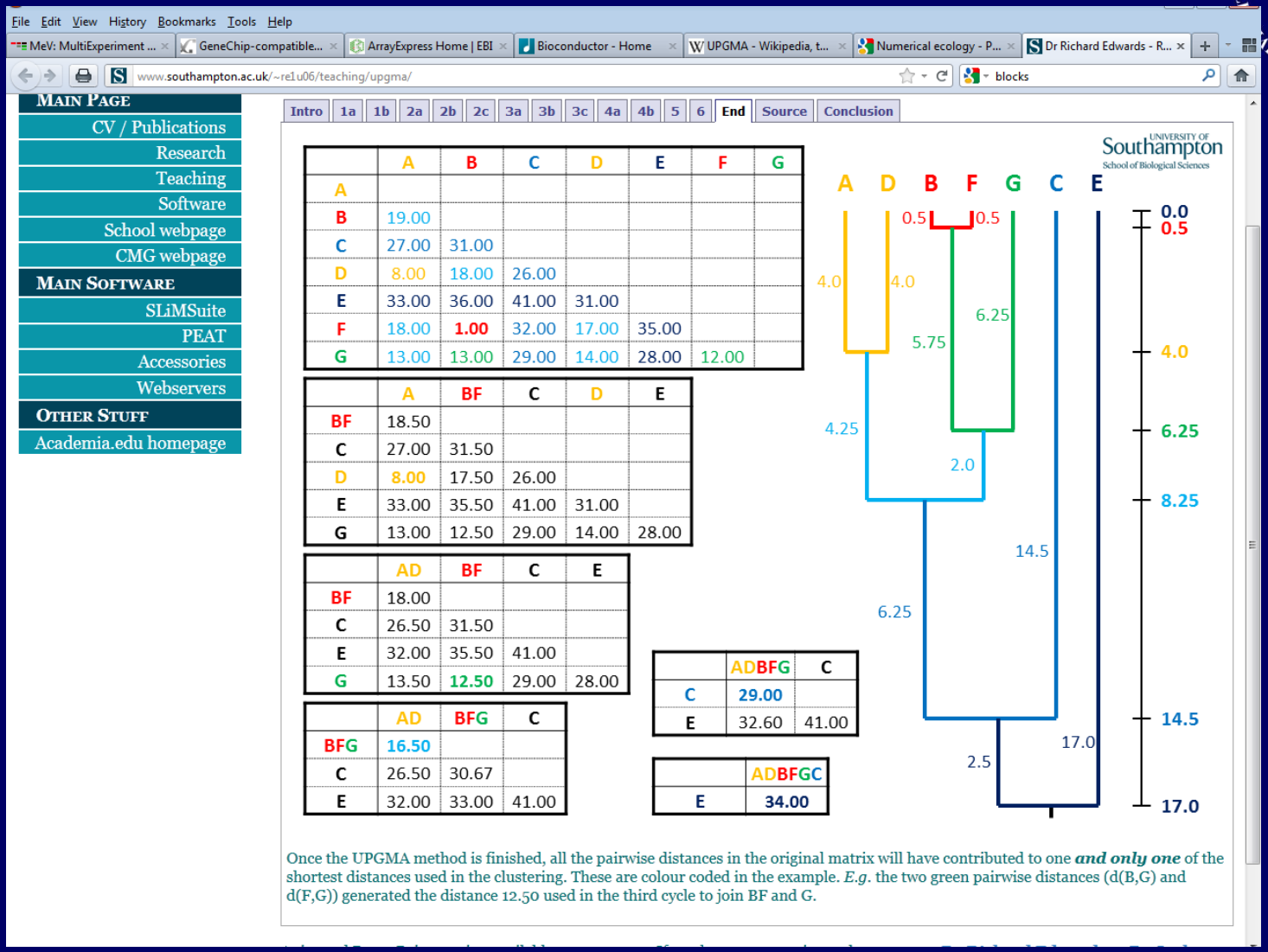
	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

	ADBFG	C	E
ADBFG			
C	29.00		
E	32.60	41.00	



	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

	ADBFGC	E
ADBFGC		
E	34.00	



File Edit View History Bookmarks Tools Help

MeV: MultiExperiment ... | GeneChip-compatible... | ArrayExpress Home | EBI | Bioconductor - Home | UPGMA - Wikipedia, t... | Numerical ecology - P... | Dr Richard Edwards - R...

www.southampton.ac.uk/~re1u06/teaching/upgma/

blocks

MAIN PAGE
CV / Publications
Research
Teaching
Software
School webpage
CMG webpage
MAIN SOFTWARE
SLiMSuite
PEAT
Accessories
Webservers
OTHER STUFF
Academia.edu homepage

Intro | 1a | 1b | 2a | 2b | 2c | 3a | 3b | 3c | 4a | 4b | 5 | 6 | End | Source | Conclusion

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

	A	BF	C	D	E
BF	18.50				
C	27.00	31.50			
D	8.00	17.50	26.00		
E	33.00	35.50	41.00	31.00	
G	13.00	12.50	29.00	14.00	28.00

	AD	BF	C	E
BF	18.00			
C	26.50	31.50		
E	32.00	35.50	41.00	
G	13.50	12.50	29.00	28.00

	AD	BFG	C
BFG	16.50		
C	26.50	30.67	
E	32.00	33.00	41.00

	ADBFG	C
C	29.00	
E	32.60	41.00

	ADBFGC
E	34.00

Once the UPGMA method is finished, all the pairwise distances in the original matrix will have contributed to one ***and only one*** of the shortest distances used in the clustering. These are colour coded in the example. *E.g.* the two green pairwise distances (d(B,G) and d(F,G)) generated the distance 12.50 used in the third cycle to join BF and G.

File Edit View History Bookmarks Tools Help

MeV: MultiExperiment ... | GeneChip-compatible... | ArrayExpress Home | EBI | Bioconductor - Home | UPGMA - Wikipedia, t... | Numerical ecology - P... | Dr Richard Edwards - R...

www.southampton.ac.uk/~re1u06/teaching/upgma/

MAIN PAGE
CV / Publications
Research
Teaching
Software
School webpage
CMG webpage
MAIN SOFTWARE
SLiMSuite
PEAT
Accessories
Webservers
OTHER STUFF
Academia.edu homepage

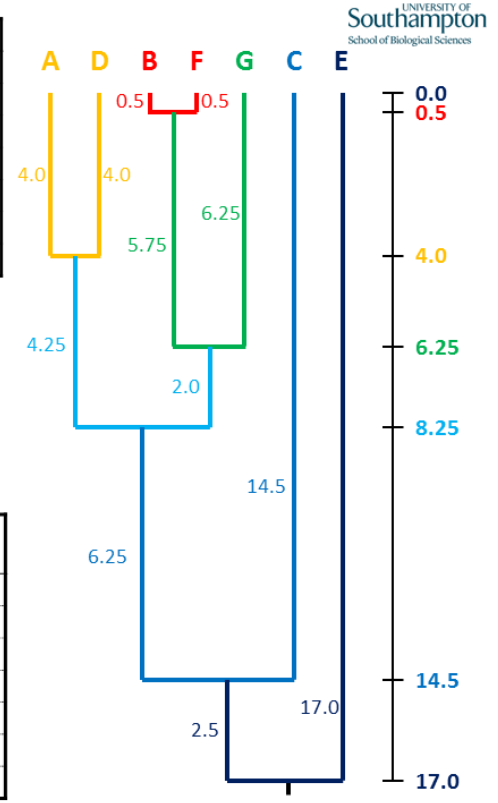
Intro | 1a | 1b | 2a | 2b | 2c | 3a | 3b | 3c | 4a | 4b | 5 | 6 | End | Source | Conclusion

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

The source data for this worked example is a selection of Cytochrome C distances from Table 3 of one of the seminal phylogenetic papers: Fitch WM & Margoliash E (1967). Construction of phylogenetic trees. *Science* **155**:279-84.

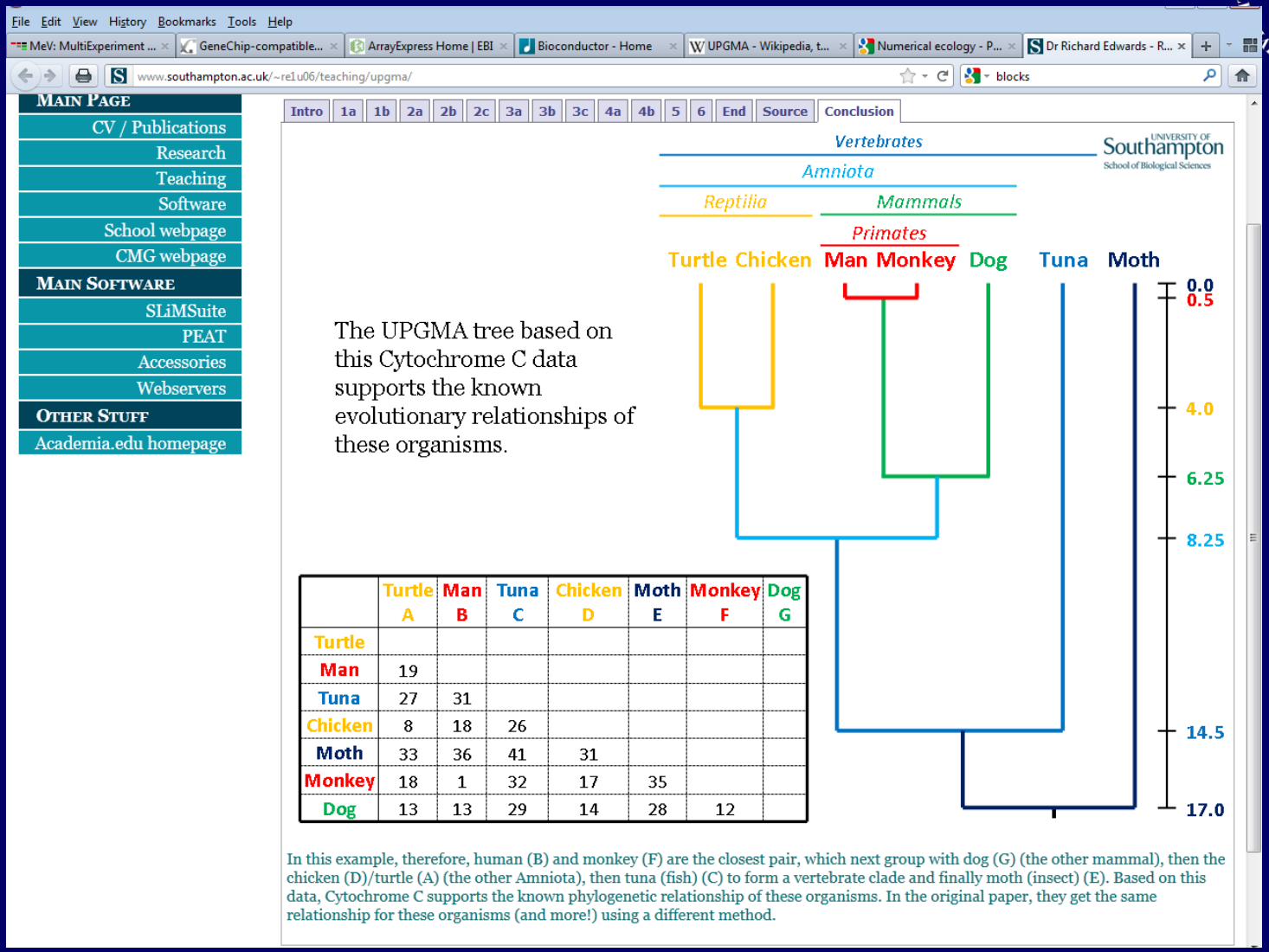
http://www.ncbi.nlm.nih.gov/pubmed/5334057

	Turtle A	Man B	Tuna C	Chicken D	Moth E	Monkey F	Dog G
Turtle							
Man	19						
Tuna	27	31					
Chicken	8	18	26				
Moth	33	36	41	31			
Monkey	18	1	32	17	35		
Dog	13	13	29	14	28	12	



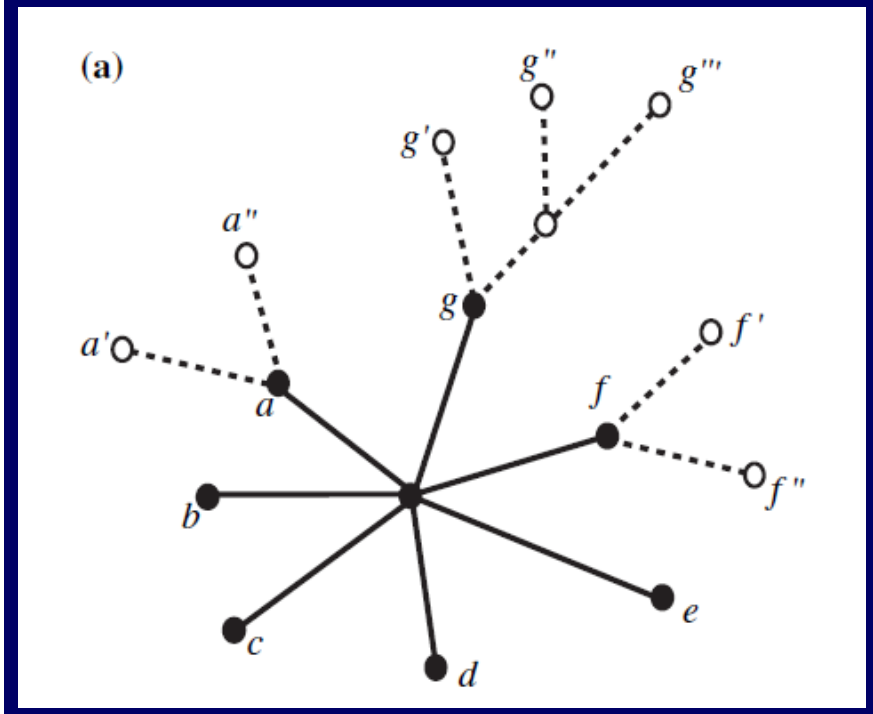
For clarity, this data represents only a subset of the taxa included in the original **Fitch & Margoliash paper**.

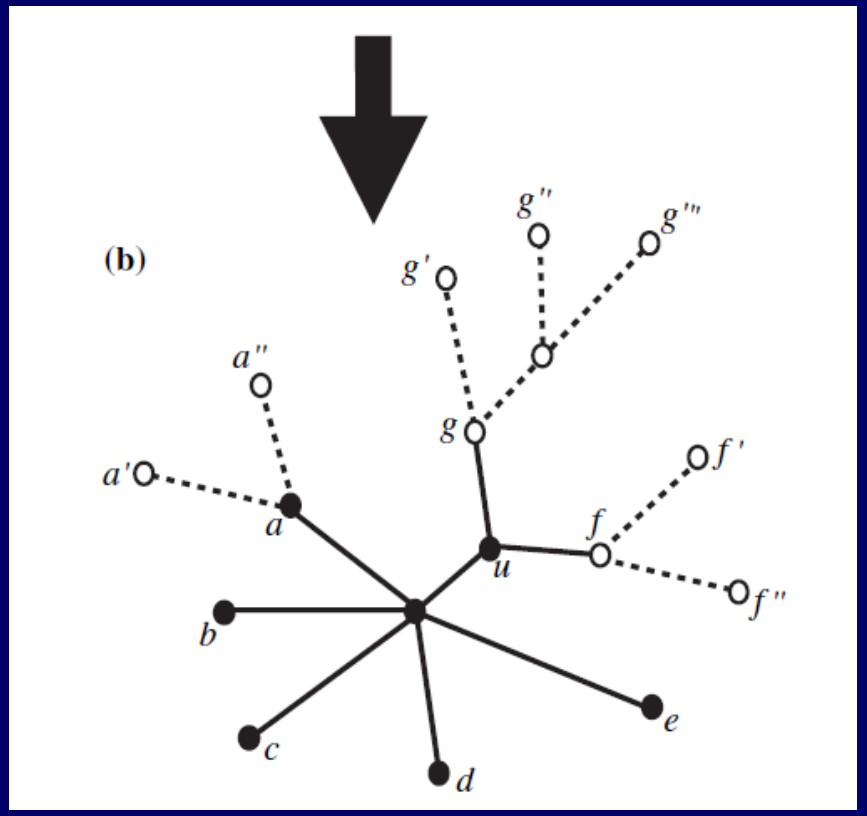
Animated PowerPoint version available upon request. If you have any questions, please contact **Dr Richard Edwards** or **Dr Joel Parker**. *This page and accompanying resources are under continual revision and development. Please report any obvious*



	Turtle A	Man B	Tuna C	Chicken D	Moth E	Monkey F	Dog G
Turtle							
Man	19						
Tuna	27	31					
Chicken	8	18	26				
Moth	33	36	41	31			
Monkey	18	1	32	17	35		
Dog	13	13	29	14	28	12	

Neighbor-Joining





Az algoritmus leírása

- Kiindulunk a távolságmátrixból
- Kezdetben a pontok csillag alakban vannak elrendezve
- Kiválasztjuk azt a két pontot, amelyek legközelebb vannak egymáshoz a többi ponthoz képest
- Új pontot hozunk létre a két pont közt, figyelembe véve a két pont távolságát egymáshoz és az összes többi ponthoz
- Helyettesítjük a két régi pont távolságát a többi ponthoz képest a távolságmátrixban
- Új ciklust kezdünk
- Az így kapott fa ágainak teljes hossza minimális lesz

Távolság

Összes pont távolsága ‘i’-től

‘i’ és ‘j’ pont

$$Q(i,j) = (r-2)d(i,j) - \sum_{k=1}^{r} d(i,k) - \sum_{k=1}^{r} d(j,k)$$

Pontok száma

Ez legyen minimális!

Összes pont távolsága ‘j’-től

Összes pont távolsága ‘f’-től

Az új pont távolsága ‘f’-től

$$d(f,u)=\frac{1}{2}d(f,g)+\frac{1}{2(r-2)}\left[\sum_{k=1}^{r}d(f,k)-\sum_{k=1}^{r}d(g,k)\right]$$

A két régi pont távolsága egymástól

Összes pont távolsága ‘g’-től

A pont eredeti távolsága ‘f’-től

Az új pont távolsága ‘f’-től

$$d(u,k)=\frac{1}{2}[d(f,k)-d(f,u)]+\frac{1}{2}[d(g,k)-d(g,u)]$$

A pont eredeti távolsága ‘g’-től

A többi pont távolsága az új ponttől

Az új pont távolsága ‘g’-től

UPGMA példa-mátrix

	A	B	C	D	E	F	G
A	0	19	27	8	33	18	13
B	19	0	31	18	36	1	13
C	27	31	0	26	41	32	29
D	8	18	26	0	31	17	14
E	33	36	41	31	0	35	28
F	18	1	32	17	35	0	12
G	13	13	29	14	28	12	0

Ág	Q min	Táv 1	Táv 2
BF	-228	0.8	0.2

2. Iterációs lépés

	A	BF	C	D	E	G
A	0	18	27	8	33	13
BF	18	0	31	17	35	12
C	27	31	0	26	41	29
D	8	17	26	0	31	14
E	33	35	41	31	0	28
G	13	12	29	14	28	0

Ág	Q min	Táv 1	Táv 2
BF	-228	0.8	0.2
AD	-163	4.375	3.625

3. Iterációs lépés

	AD	BF	C	E	G
AD	0	13.5	22.5	28	9.5
BF	13.5	0	31	35	12
C	22.5	31	0	41	29
E	28	35	41	0	28
G	9.5	12	29	28	0

Ág	Q min	Táv 1	Táv 2
BF	-228	0.8	0.2
AD	-163	4.375	3.625
BFG	-134	8.16667	3.83333

4. Iterációs lépés

	AD	BFG	C	E
AD	0	5.5	22.5	28
BFG	5.5	0	24	25.5
C	22.5	24	0	41
E	28	25.5	41	0

Ág	Q min	Táv 1	Táv 2
BF	-228	0.8	0.2
AD	-163	4.375	3.625
BFG	-134	8.16667	3.83333
ADBFG	-100	3	2.5

5. Iterációs lépés

	ADBFG	C	E
ADBFG	0	22.5	28
C	22.5	0	41
E	28	41	0

Ág	Q min	Táv 1	Táv 2
BF	-228	0.8	0.2
AD	-163	4.375	3.625
BFG	-134	8.16667	3.83333
ADBFG	-100	3	2.5
ADBFGC	-85.5	1.75	18.75

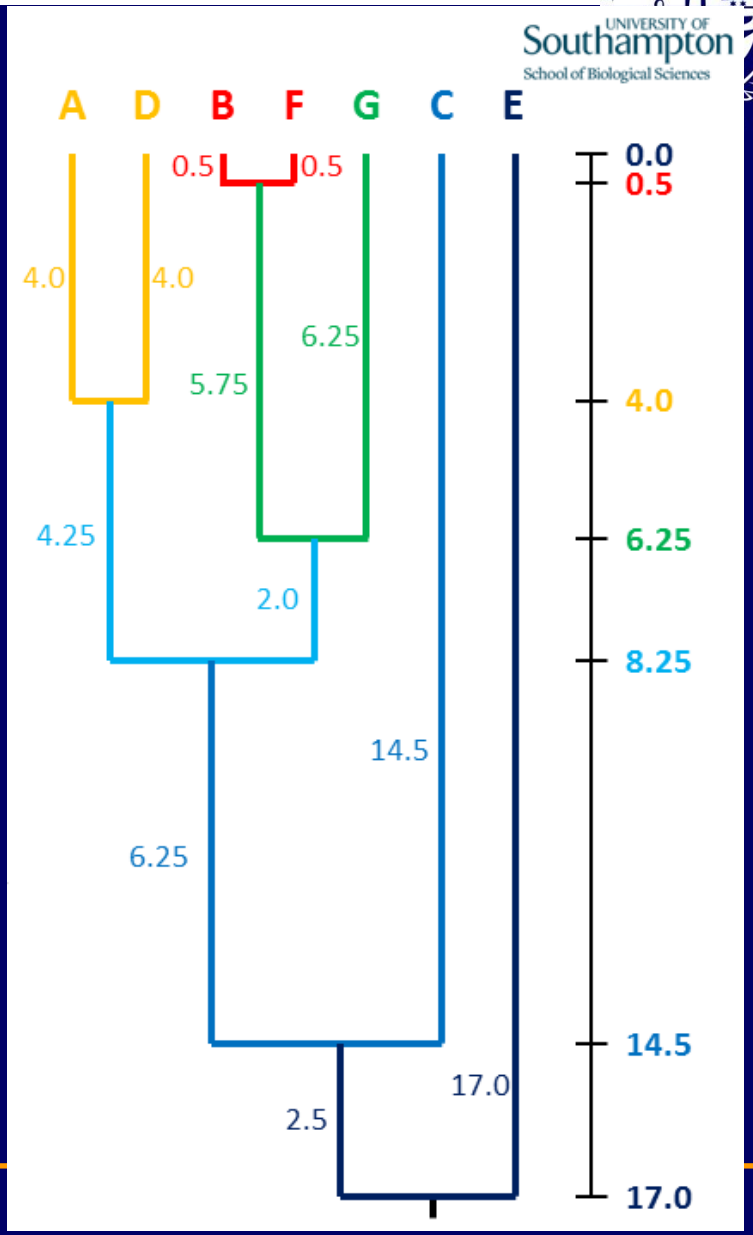
6. Iterációs lépés

	ADBFGC	E
ADBFGC	0	22.25
E	22.25	0

Ág	Q min	Táv 1	Táv 2
BF	-228	0.8	0.2
AD	-163	4.375	3.625
BFG	-134	8.16667	3.83333
ADBFG	-100	3	2.5
ADBFGC	-85.5	1.75	18.75

NJ vs. UPGMA

Ág	Táv 1	Táv 2
BF	0.8	0.2
AD	4.375	3.625
BFG	8.16667	3.83333
ADBFG	3	2.5
ADBFGC	1.75	18.75
ADBFGCE	0	22.25

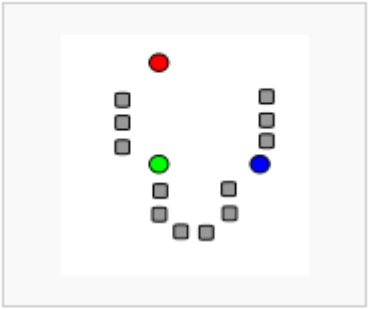


K-közép klaszterezés

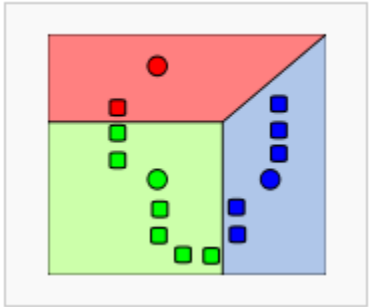
- Keressük *n* pont felosztását *k* klaszterbe
- A klaszterelet az őket alkotó pontok súlypontjával jellemezzük
- A probléma analitikusan nem megoldható
- Heurisztikus megoldás van – iteráció
- Nem garantált az optimális eredmény
- Ismételt futtatás után össze kell vetni a kapott klasztereket

Az algoritmus

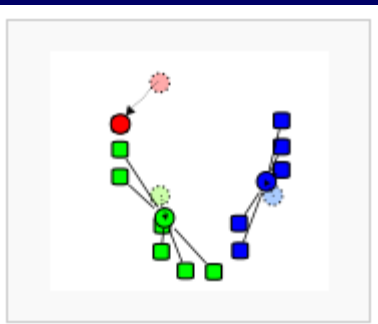
- Két lépéses eljárás:
  - Kiszámoljuk a klaszterek súlypontját adott felosztás mellett
  - Átsoroljuk a pontokat, ha van közelebbi klaszter az eredeti besoroláshoz képest
- Előre tudni kell, hány klaszterünk lesz



1) *k* initial "means" (in this case *k*=3) are randomly selected from the data set (shown in color).



2) *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

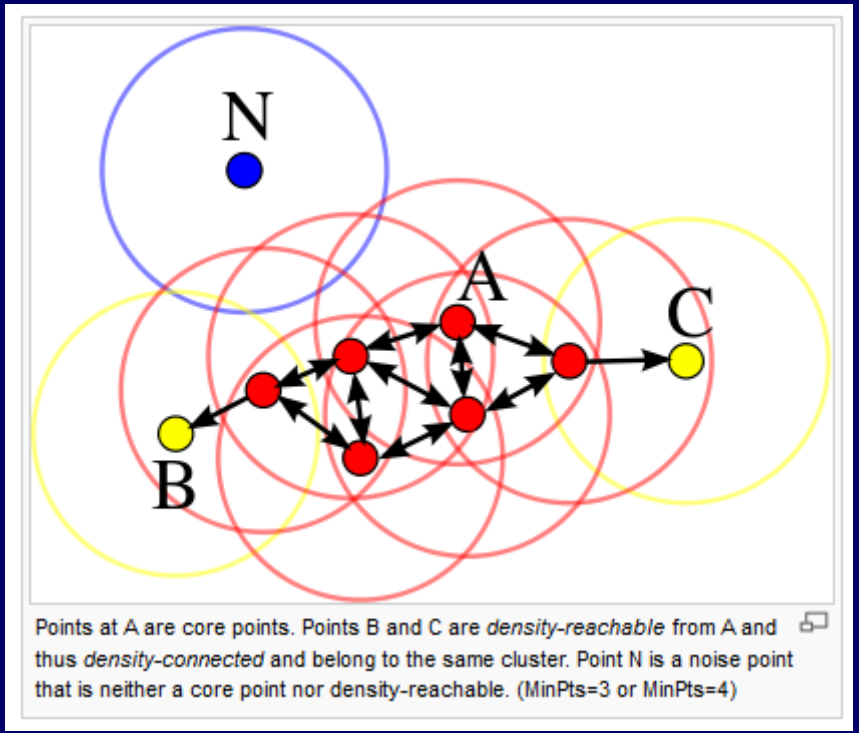


3) The centroid of each of the *k* clusters becomes the new means.

4) Steps 2 and 3 are repeated until convergence has been reached.

A DBSCAN algoritmus

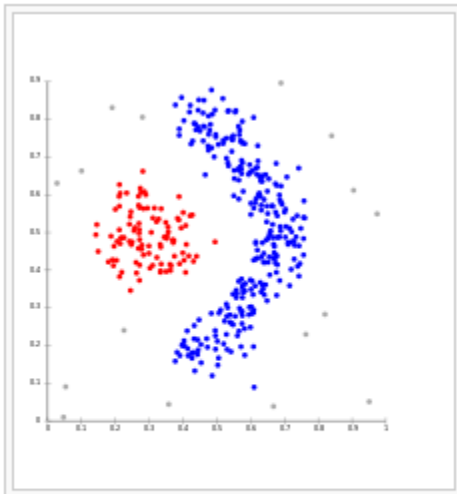
- A módszer 'sűrűség' alapú
- Két bemeneti paramétere van:
  - Minimális távolság
  - Szomszédos pontok száma
- Lehet kiszóró pont – 'zaj'



Points at A are core points. Points B and C are *density-reachable* from A and thus *density-connected* and belong to the same cluster. Point N is a noise point that is neither a core point nor density-reachable. (MinPts=3 or MinPts=4)

A módszer tulajdonságai

- Nem kell tudni előre a klaszterek számát
- A klaszterekeken belül nincs rendszer
- Eltérő sűrűségű klaszterek kezelése nehéz
- A módszer gyors, de sok memória kell



DBSCAN can find non-linearly separable clusters. This dataset cannot be adequately clustered with k-means or EM clustering.

Klaszterezés a biológiában

Filogenetika

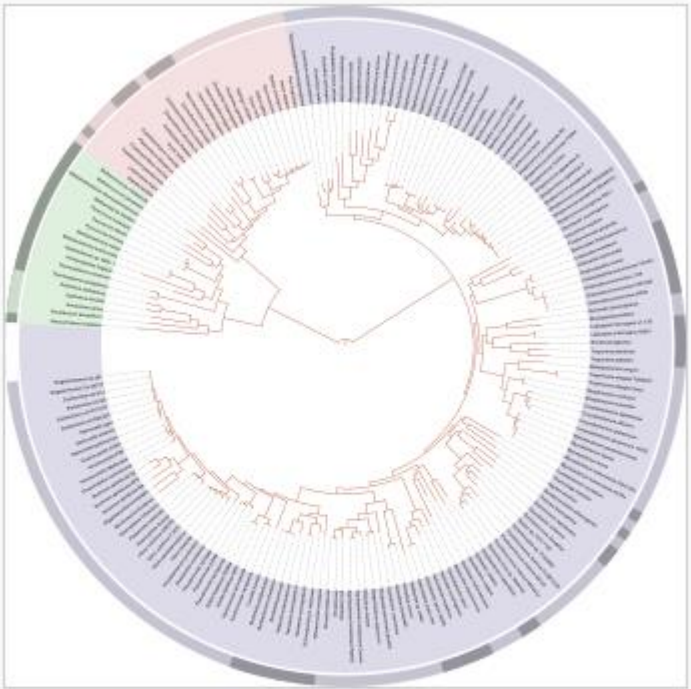
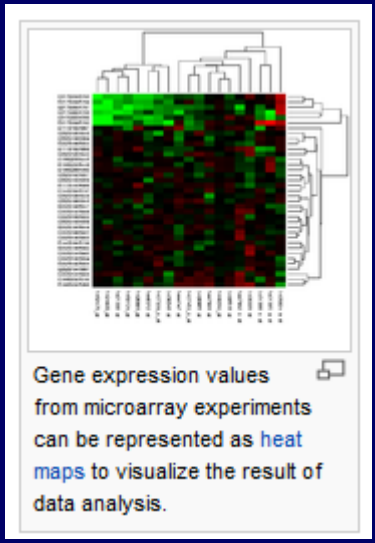


Fig. 3: A highly resolved, automatically generated Tree Of Life, based on completely sequenced genomes.[4][5]

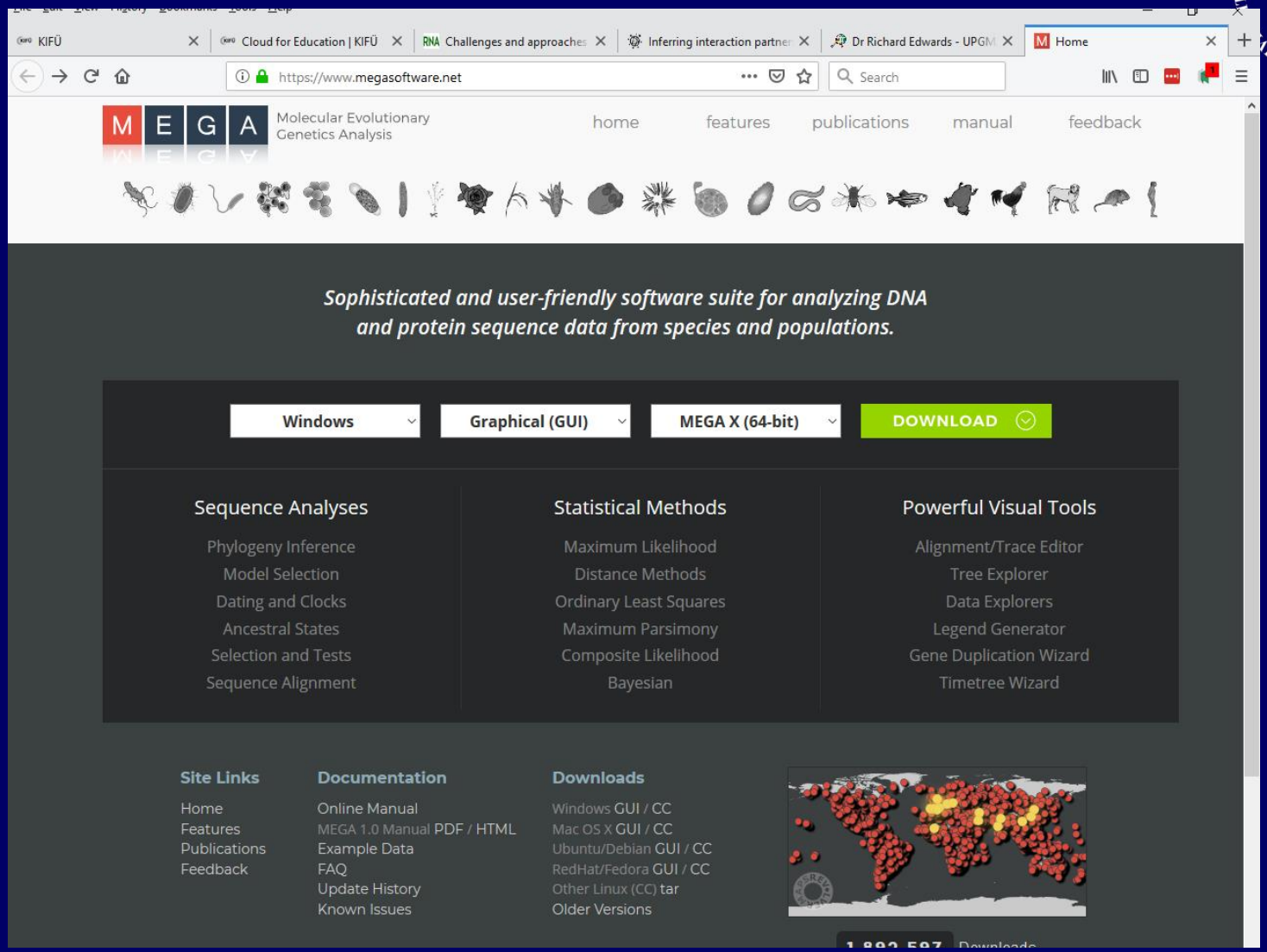
Gén-chip analízis

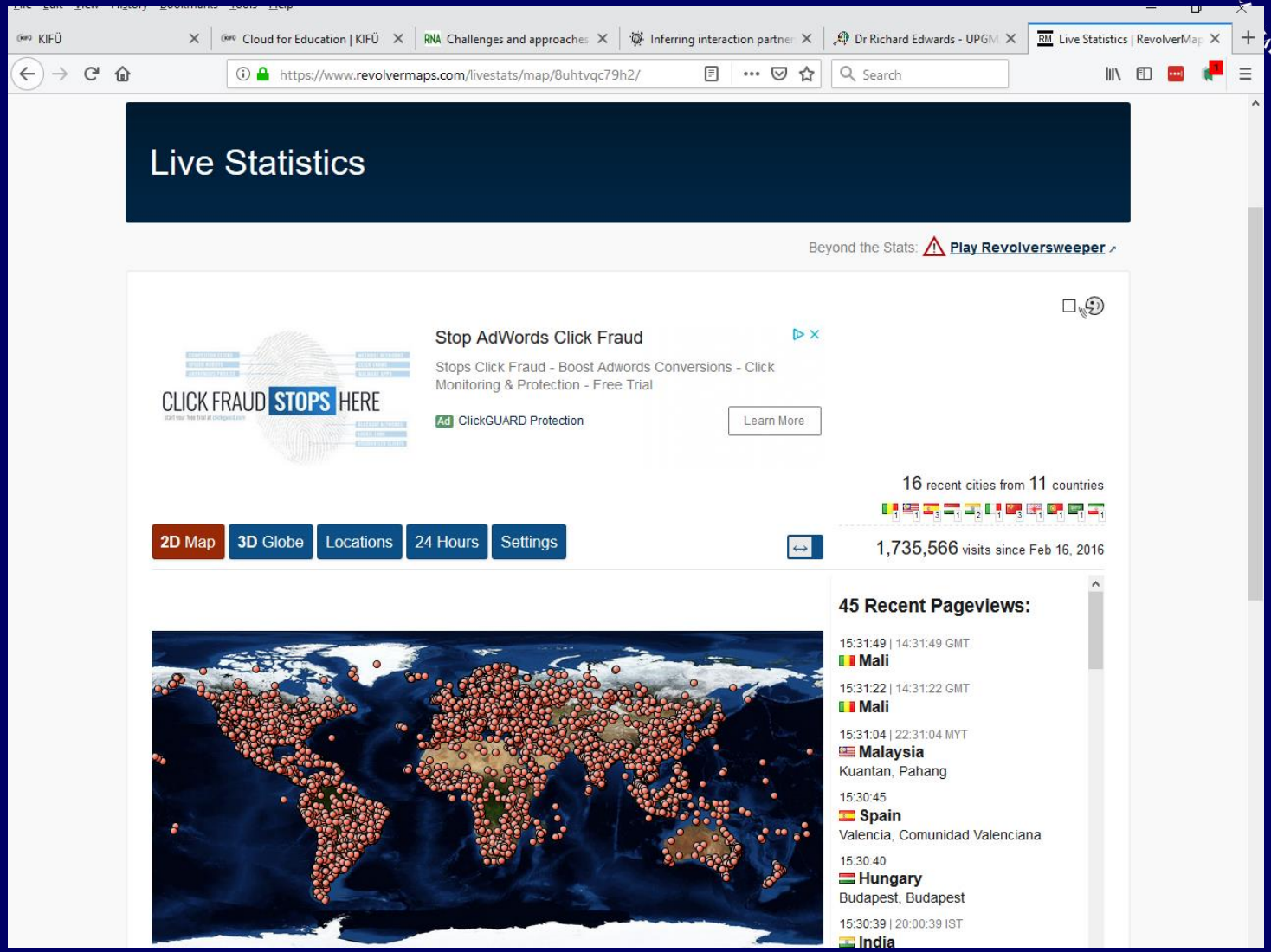


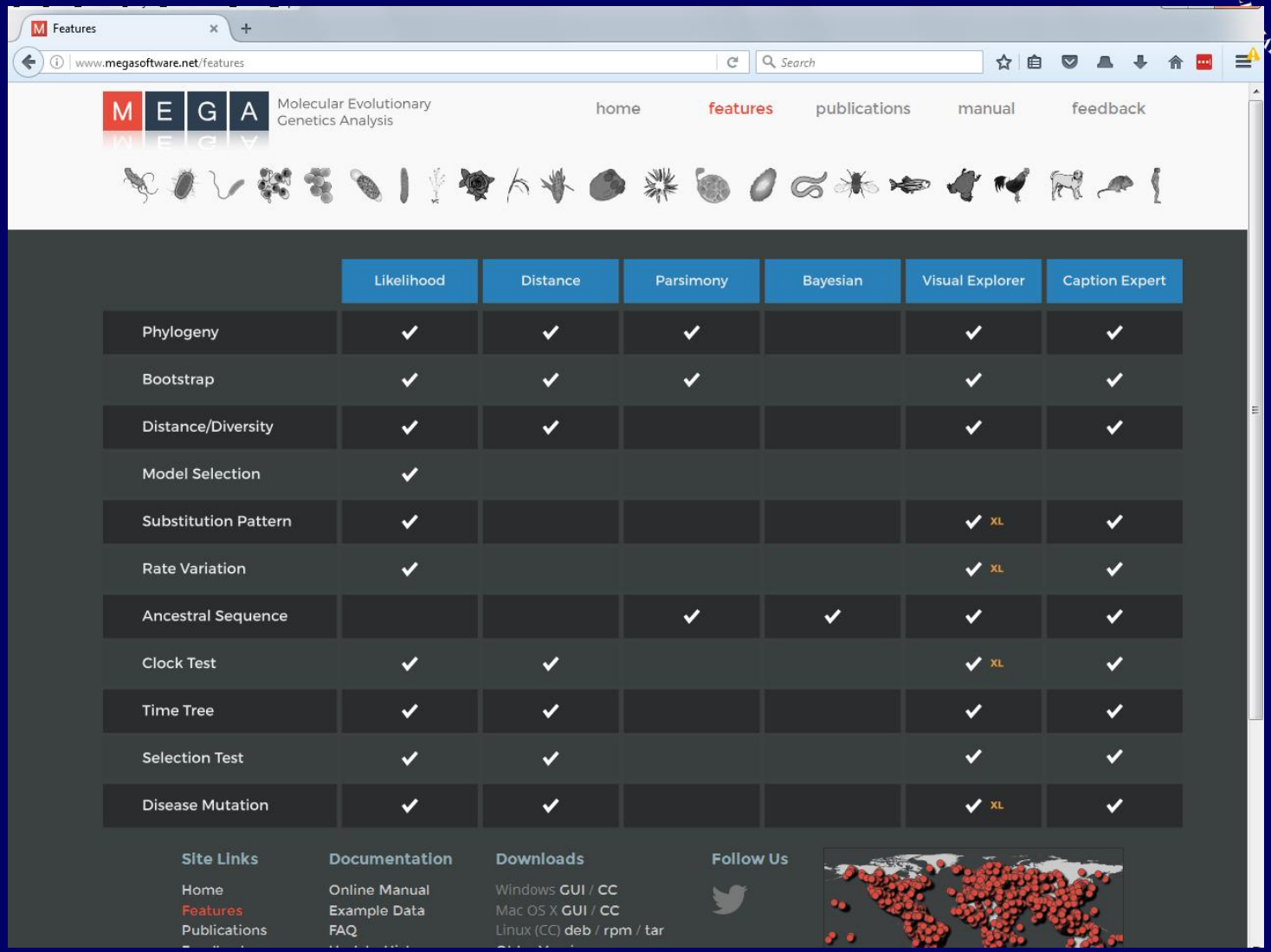
Gene expression values from microarray experiments can be represented as heat maps to visualize the result of data analysis.

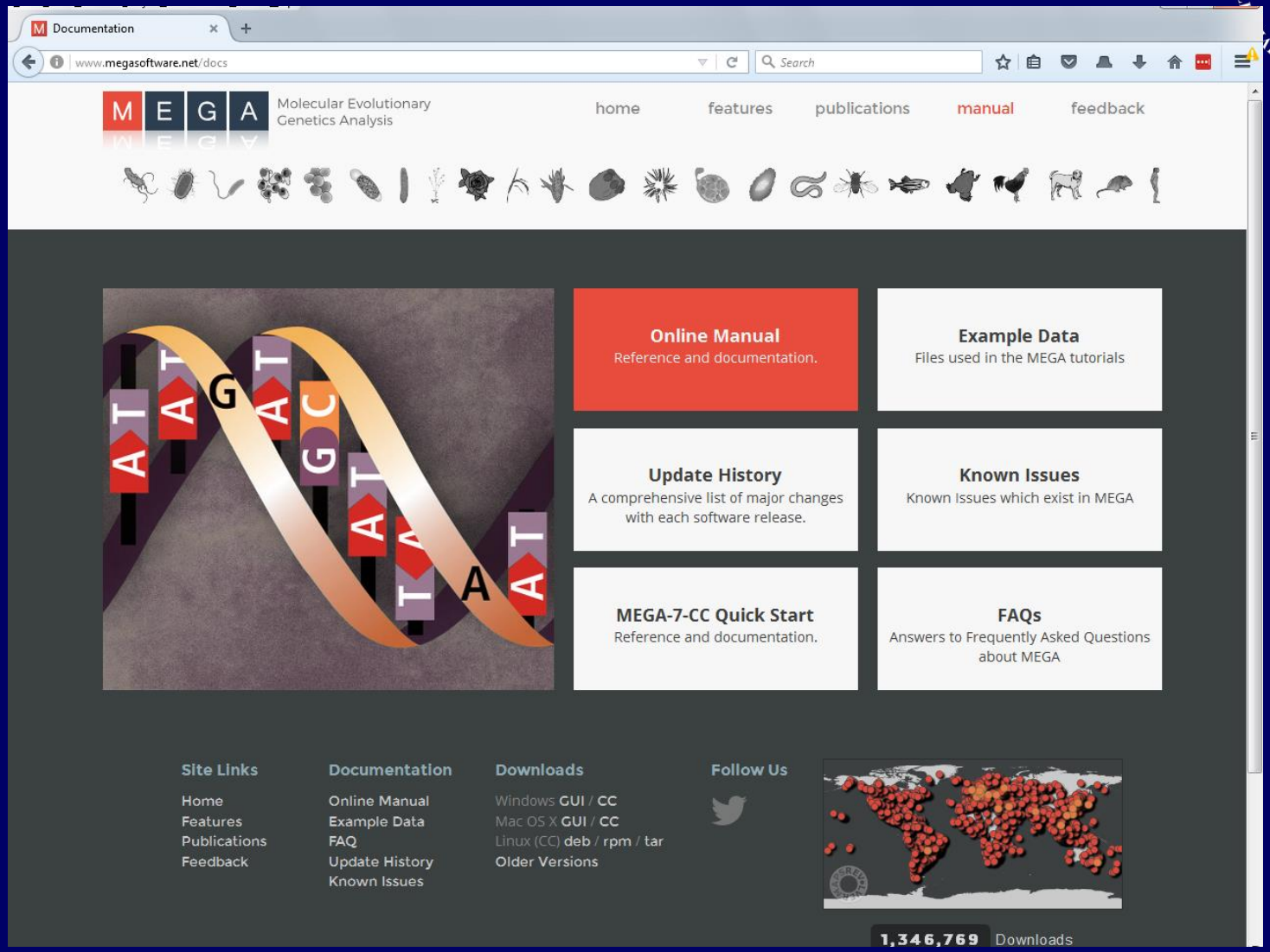
Klaszterező alkalmazás

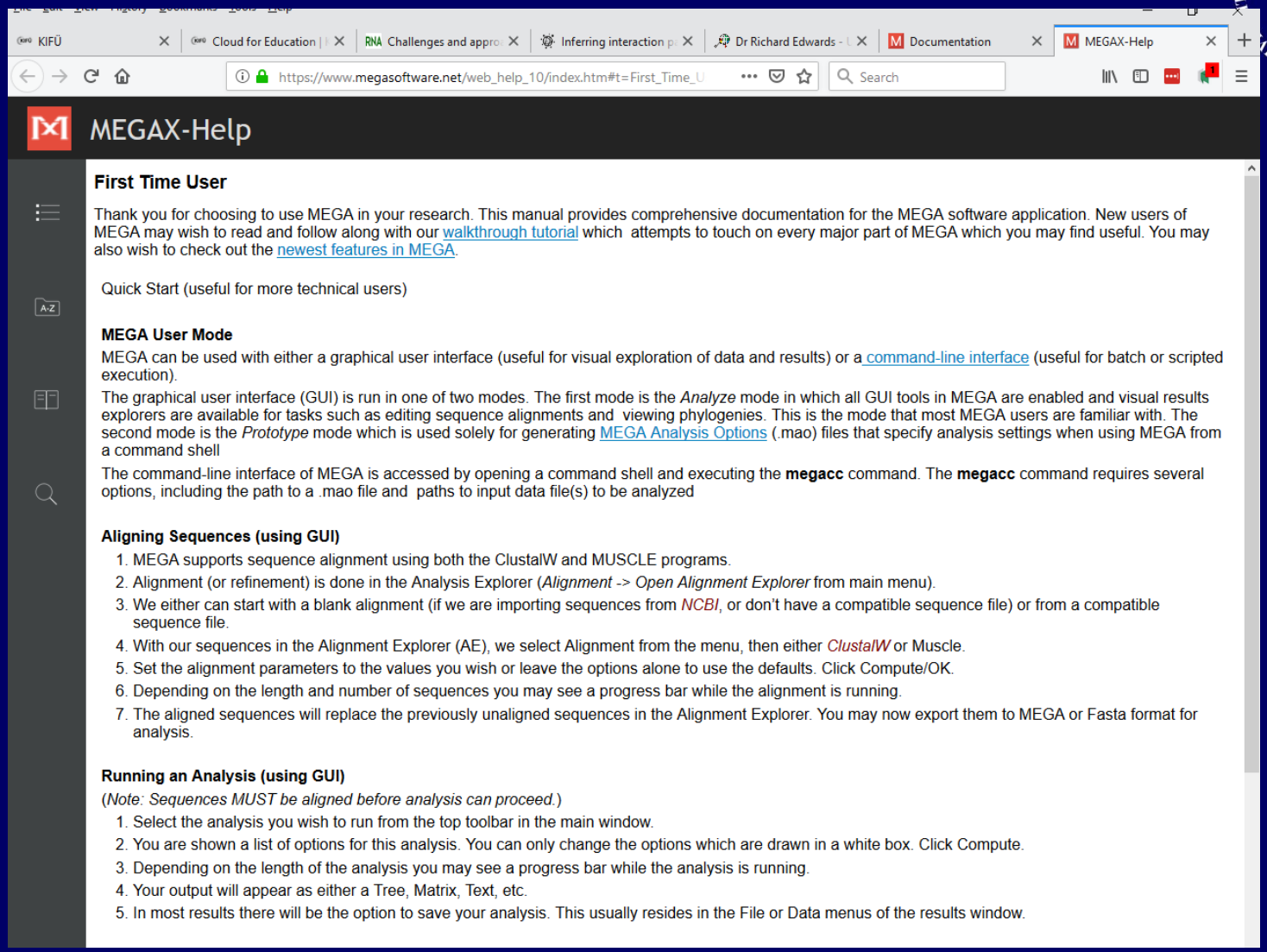
- "Molecular evolutionary genetic analysis"
- MEGA: http://www.megasoftware.net/
- Ingyenesen elérhető akadémiai felhasználásra
- Widows és Mac változat is

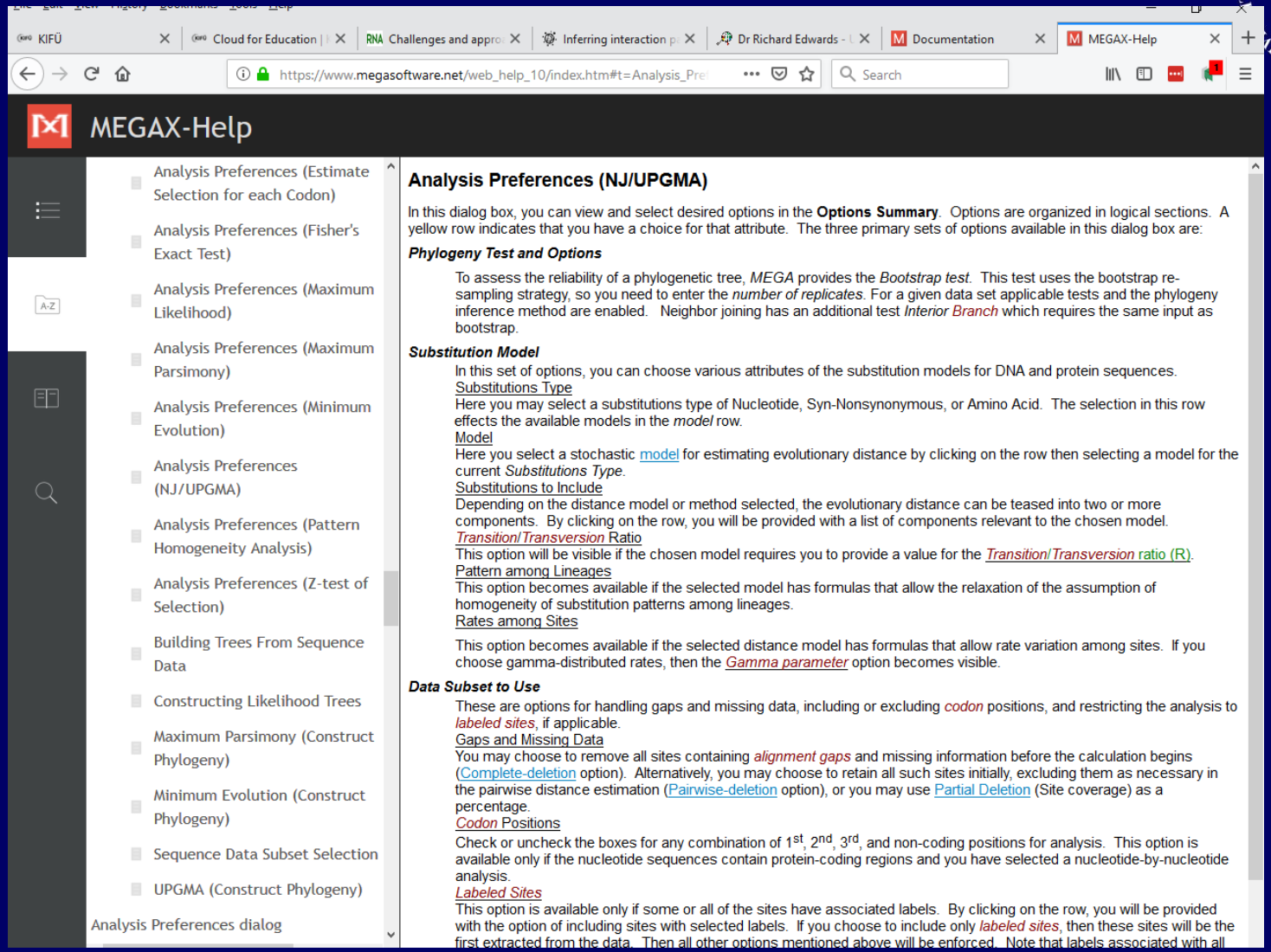

KIFÜ
Cloud for Education | KIFÜ
Challenges and approaches
Inferring interaction partner
Dr Richard Edwards - UPGM
Home
https://www.megasoftware.net
Search
MEGA
Molecular Evolutionary Genetics Analysis
home
features
publications
manual
feedback
Sophisticated and user-friendly software suite for analyzing DNA and protein sequence data from species and populations.
Windows
Graphical (GUI)
MEGA X (64-bit)
DOWNLOAD
Sequence Analyses
Phylogeny Inference
Model Selection
Dating and Clocks
Ancestral States
Selection and Tests
Sequence Alignment
Statistical Methods
Maximum Likelihood
Distance Methods
Ordinary Least Squares
Maximum Parsimony
Composite Likelihood
Bayesian
Powerful Visual Tools
Alignment/Trace Editor
Tree Explorer
Data Explorers
Legend Generator
Gene Duplication Wizard
Timetree Wizard
Site Links
Home
Features
Publications
Feedback
Documentation
Online Manual
MEGA 1.0 Manual PDF / HTML
Example Data
FAQ
Update History
Known Issues
Downloads
Windows GUI / CC
Mac OS X GUI / CC
Ubuntu/Debian GUI / CC
RedHat/Fedora GUI / CC
Other Linux (CC) tar
Older Versions


KIFÜ
Cloud for Education | KIFÜ
Challenges and approaches
Inferring interaction partner
Dr Richard Edwards - UPGM
Live Statistics | RevolverMap
https://www.revolvermaps.com/livestats/map/8uhtvqc79h2/
Search
Live Statistics
Beyond the Stats: Play Revolversweeper
Stop AdWords Click Fraud
Stops Click Fraud - Boost Adwords Conversions - Click Monitoring & Protection - Free Trial
CLICK FRAUD STOPS HERE
Ad ClickGUARD Protection
Learn More
16 recent cities from 11 countries
2D Map
3D Globe
Locations
24 Hours
Settings
1,735,566 visits since Feb 16, 2016
45 Recent Pageviews:
15:31:49 | 14:31:49 GMT
Mali
15:31:22 | 14:31:22 GMT
Mali
15:31:04 | 22:31:04 MYT
Malaysia
Kuantan, Pahang
15:30:45
Spain
Valencia, Comunidad Valenciana
15:30:40
Hungary
Budapest, Budapest
15:30:39 | 20:00:39 IST
India


Features
www.megasoftware.net/features
Search
MEGA
Molecular Evolutionary Genetics Analysis
home
features
publications
manual
feedback
Likelihood
Distance
Parsimony
Bayesian
Visual Explorer
Caption Expert
Phylogeny
Bootstrap
Distance/Diversity
Model Selection
Substitution Pattern
Rate Variation
Ancestral Sequence
Clock Test
Time Tree
Selection Test
Disease Mutation
XL
Site Links
Home
Features
Publications
Documentation
Online Manual
Example Data
FAQ
Downloads
Windows GUI / CC
Mac OS X GUI / CC
Linux (CC) deb / rpm / tar
Follow Us


Documentation
www.megasoftware.net/docs
Search
M E G A
Molecular Evolutionary Genetics Analysis
home
features
publications
manual
feedback
Online Manual
Reference and documentation.
Example Data
Files used in the MEGA tutorials
Update History
A comprehensive list of major changes with each software release.
Known Issues
Known Issues which exist in MEGA
MEGA-7-CC Quick Start
Reference and documentation.
FAQs
Answers to Frequently Asked Questions about MEGA
Site Links
Home
Features
Publications
Feedback
Documentation
Online Manual
Example Data
FAQ
Update History
Known Issues
Downloads
Windows GUI / CC
Mac OS X GUI / CC
Linux (CC) deb / rpm / tar
Older Versions
Follow Us
1,346,769 Downloads


KIFÜ
Cloud for Education
Challenges and appro
Inferring interaction p
Dr Richard Edwards -
Documentation
MEGAX-Help
https://www.megasoftware.net/web_help_10/index.htm#t=First_Time_U
Search
MEGAX-Help
First Time User
Thank you for choosing to use MEGA in your research. This manual provides comprehensive documentation for the MEGA software application. New users of MEGA may wish to read and follow along with our walkthrough tutorial which attempts to touch on every major part of MEGA which you may find useful. You may also wish to check out the newest features in MEGA.
Quick Start (useful for more technical users)
MEGA User Mode
MEGA can be used with either a graphical user interface (useful for visual exploration of data and results) or a command-line interface (useful for batch or scripted execution).
The graphical user interface (GUI) is run in one of two modes. The first mode is the Analyze mode in which all GUI tools in MEGA are enabled and visual results explorers are available for tasks such as editing sequence alignments and viewing phylogenies. This is the mode that most MEGA users are familiar with. The second mode is the Prototype mode which is used solely for generating MEGA Analysis Options (.mao) files that specify analysis settings when using MEGA from a command shell
The command-line interface of MEGA is accessed by opening a command shell and executing the megacc command. The megacc command requires several options, including the path to a .mao file and paths to input data file(s) to be analyzed
Aligning Sequences (using GUI)
1. MEGA supports sequence alignment using both the ClustalW and MUSCLE programs.
2. Alignment (or refinement) is done in the Analysis Explorer (Alignment -> Open Alignment Explorer from main menu).
3. We either can start with a blank alignment (if we are importing sequences from NCBI, or don't have a compatible sequence file) or from a compatible sequence file.
4. With our sequences in the Alignment Explorer (AE), we select Alignment from the menu, then either ClustalW or Muscle.
5. Set the alignment parameters to the values you wish or leave the options alone to use the defaults. Click Compute/OK.
6. Depending on the length and number of sequences you may see a progress bar while the alignment is running.
7. The aligned sequences will replace the previously unaligned sequences in the Alignment Explorer. You may now export them to MEGA or Fasta format for analysis.
Running an Analysis (using GUI)
(Note: Sequences MUST be aligned before analysis can proceed.)
1. Select the analysis you wish to run from the top toolbar in the main window.
2. You are shown a list of options for this analysis. You can only change the options which are drawn in a white box. Click Compute.
3. Depending on the length of the analysis you may see a progress bar while the analysis is running.
4. Your output will appear as either a Tree, Matrix, Text, etc.
5. In most results there will be the option to save your analysis. This usually resides in the File or Data menus of the results window.


KIFÜ
Cloud for Education
Challenges and appro
Inferring interaction p
Dr Richard Edwards -
Documentation
MEGAX-Help
https://www.megasoftware.net/web_help_10/index.htm#t=Analysis_Pre
Search
MEGAX-Help
Analysis Preferences (Estimate Selection for each Codon)
Analysis Preferences (Fisher's Exact Test)
Analysis Preferences (Maximum Likelihood)
Analysis Preferences (Maximum Parsimony)
Analysis Preferences (Minimum Evolution)
Analysis Preferences (NJ/UPGMA)
Analysis Preferences (Pattern Homogeneity Analysis)
Analysis Preferences (Z-test of Selection)
Building Trees From Sequence Data
Constructing Likelihood Trees
Maximum Parsimony (Construct Phylogeny)
Minimum Evolution (Construct Phylogeny)
Sequence Data Subset Selection
UPGMA (Construct Phylogeny)
Analysis Preferences dialog
Analysis Preferences (NJ/UPGMA)
In this dialog box, you can view and select desired options in the Options Summary. Options are organized in logical sections. A yellow row indicates that you have a choice for that attribute. The three primary sets of options available in this dialog box are:
Phylogeny Test and Options
To assess the reliability of a phylogenetic tree, MEGA provides the Bootstrap test. This test uses the bootstrap re-sampling strategy, so you need to enter the number of replicates. For a given data set applicable tests and the phylogeny inference method are enabled. Neighbor joining has an additional test Interior Branch which requires the same input as bootstrap.
Substitution Model
In this set of options, you can choose various attributes of the substitution models for DNA and protein sequences.
Substitutions Type
Here you may select a substitutions type of Nucleotide, Syn-Nonsynonymous, or Amino Acid. The selection in this row effects the available models in the model row.
Model
Here you select a stochastic model for estimating evolutionary distance by clicking on the row then selecting a model for the current Substitutions Type.
Substitutions to Include
Depending on the distance model or method selected, the evolutionary distance can be teased into two or more components. By clicking on the row, you will be provided with a list of components relevant to the chosen model.
Transition/Transversion Ratio
This option will be visible if the chosen model requires you to provide a value for the Transition/Transversion ratio (R).
Pattern among Lineages
This option becomes available if the selected model has formulas that allow the relaxation of the assumption of homogeneity of substitution patterns among lineages.
Rates among Sites
This option becomes available if the selected distance model has formulas that allow rate variation among sites. If you choose gamma-distributed rates, then the Gamma parameter option becomes visible.
Data Subset to Use
These are options for handling gaps and missing data, including or excluding codon positions, and restricting the analysis to labeled sites, if applicable.
Gaps and Missing Data
You may choose to remove all sites containing alignment gaps and missing information before the calculation begins (Complete-deletion option). Alternatively, you may choose to retain all such sites initially, excluding them as necessary in the pairwise distance estimation (Pairwise-deletion option), or you may use Partial Deletion (Site coverage) as a percentage.
Codon Positions
Check or uncheck the boxes for any combination of 1st, 2nd, 3rd, and non-coding positions for analysis. This option is available only if the nucleotide sequences contain protein-coding regions and you have selected a nucleotide-by-nucleotide analysis.
Labeled Sites
This option is available only if some or all of the sites have associated labels. By clicking on the row, you will be provided with the option of including sites with selected labels. If you choose to include only labeled sites, then these sites will be the first extracted from the data. Then all other options mentioned above will be enforced. Note that labels associated with all

Mit tanultunk ma?

- A klaszterezés eredménye függ az alkalmazott:
  - metrikától
  - klaszterező kritériumtól
- Az eredmények megjelenítése:
  - nem egyszerű
  - nem egyértelmű
- Nincs legjobb megoldás, minden feladat más

Feladat 10.

- Rendezd klaszterbe az aminósavakat!