

Upcoming Challenges for Multiple Sequence Alignment Methods


Cédric Notredame

Comparative Bioinformatics Group

Bioinformatics and Genomics Program



What is NGS sequencing changing for Regular Biology ?

HHMI  [HOME](#) | [SEARCH](#) | [BROWSE](#) | [FTP](#) | [HELP](#)

Pfam
keyword search

Family: GP120 (PF00516)

10 architectures 75195 sequences 3 interactions 95 species 17 structures

Summary

Envelope glycoprotein GP120

The entry of HIV requires interaction of viral GP120 with [P01730](#) and a chemokine receptor on the cell surface.

Literature references

1. Kwong PD, Wyatt R, Robinson J, Sweet RW, Sodroski J, Hendrickson WA; , Nature 1998;393:648-659.: Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. [PUBMED:9641677](#)

Interpro entry [IPR000777](#)

The entry of HIV requires interaction of viral GP120, an envelope glycoprotein with human T-cell surface glycoprotein CD4 and a chemokine receptor on the cell surface. These envelope glycoproteins are found in HIV types 1 and 2, and Simian Immunodeficiency virus (SIV).

Gene Ontology

Cellular component [viral envelope \(GO:0019031\)](#)

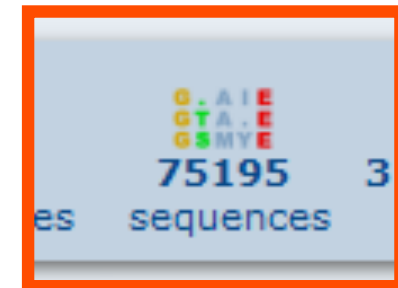
Internal database links

SCOOP: [APG12](#)

Jump to...

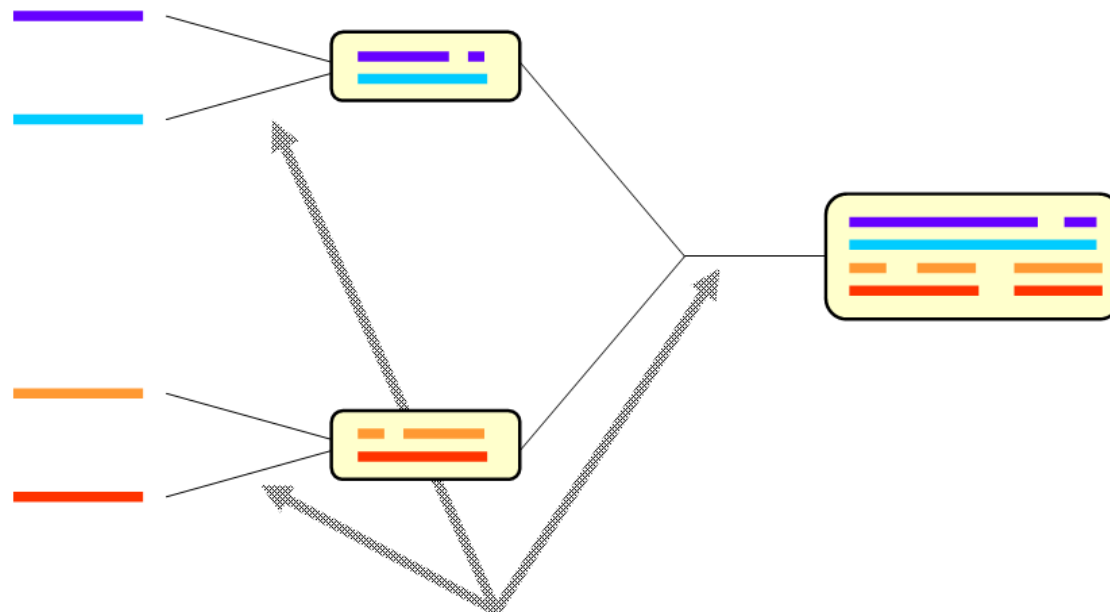
enter ID/acc

Domain organisation
Alignments
HMM logo
Trees
Curation & models
Species
Interactions
Structures



Aligning Very Large Datasets is Challenging

Feng and Dolittle, 1980; Taylor 1981



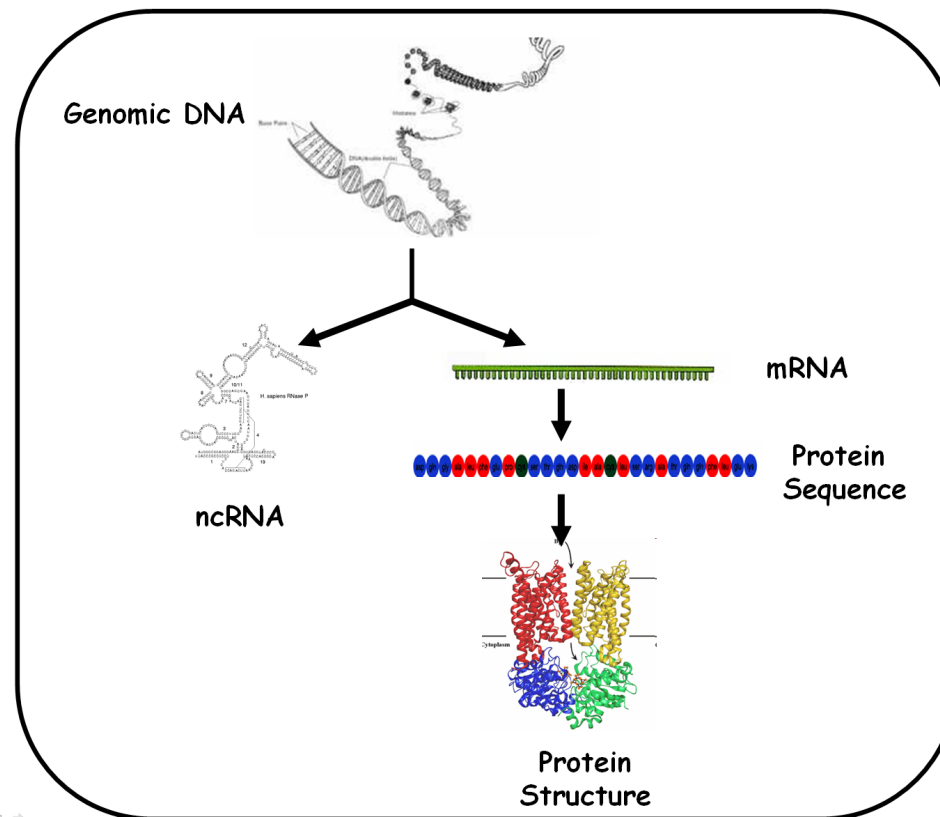
Dynamic Programming Using A Substitution Matrix

Ref. 20_BG.ES.99.R77.AY586544	TGTGGAAAAGGAGGGACATCAAATGAAAAGACTGCACA-----GAGAG
Ref. 21_A2D.KE.91.KNH1254.AY945737	TGTGGAAAAGGAAAGGGACCCAAATGAAAAGATTGTACG-----GAGAG
Ref. 21_A2D.KE.99.KER2003.AF457051	TGTGGAAAAGGAAAGGGACCCAAATGAGAGATTGCACG-----GAAAG
Ref. 22_01A1.CM.01.01CM_0001BBY.AY371159	TGTGGGAAAAGGAAAGGACACCCAAATGAAAAGACTGCACTCTTACTCTTGAGAG
Ref. 23_BG.CU.03.CB118.AY900571	TGTGGAAAAGGATGGACATCAAATGAAAAGACTGCACA-----GAAGGGAG
Ref. 23_BG.CU.03.CB347.AY900572	TGTGGAAAAGGAGGGACATCAAATGAAAAGACTGCACA-----GAGAG
Ref. 24_BG.CU.03.CB378.AY900574	TGTGGAAAAGGAGGGACATCAAATGAAAAGACTGCACA-----GAGAG
Ref. 24_BG.CU.03.CB471.AY900575	TGTGGAAAAGGAGGGACATCAAATGAAAAGACTGCACA-----GAGAG
Ref. 25_cpx.CM.06.06CM_BA_040.EU693240	TGTGGGAAAAGAAGGACATCAGATGAAAAGACTGCACA-----GAGAG
Ref. 25_cpx.SA.03.J11233.EU697906	TGTGGAAAAGGAGGGACATCAAATGAAAAGACTGCACG-----GAGAG
Ref. 25_cpx.SA.03.J11451.EU697908	TGTGGGAAAAGGAGGGACATCAAATGAAAAGACTGCACR-----GARAG
Ref. 27_cpx.CD.97.97CDKTB49.AJ404325	TGTGGAAAAGGAGGGACATCAAATGAAAAGACTGTACA-----GAGAG
Ref. 27_cpx.FR.04.04CD_FR_KZS.AM851091	TGTGGAAAAGAGAGGGACATCAAATGAAAAGACTGTACA-----GAGAG
Ref. 28_BF.BR.99.BREPM12313.DQ085872	TGTGGAAAGAGAAGGACACCCAAATGAAAAGACTGTACT-----GAAAG
Ref. 28_BF.BR.99.BREPM12609.DQ085873	TGTGGAAAGAGAAGGACACCCAAATGAAAAGATTGCACT-----GAAAG
Ref. 28_BF.BR.99.BREPM12817.DQ085874	TGTGGAAAAGGAAAGGACATCAAATGAAAAGACTGCACT-----GAAAG
Ref. 29_BF.BR.01.BREPM16704.DQ085876	TGTGGAAAAAGAAGGACACCCAAATGAAAAGAATGCACT-----GAAAG
Ref. 29_BF.BR.99.BREPM11948.DQ085871	TGTGGAAAGAGAAGGACACCCAAATGAAAAGACTGCACT-----GAAAG
Ref. 31_BC.BR.02.110PA.EF091932	TGTGGAAAAAGAAGGACACCCAAATGAAAAGAATGTACT-----GAGAG
Ref. 31_BC.BR.04.04BR142.AY727527	TGTGGAAAAGGAAAGGACACCCAAATGAAAAGACTGTAAT-----AATGAGAG
Ref. 32_06A1.EE.01.EE0369.AY535660	TGTGGACAGGAAGGCCATCAAATGAAAAGACTGCACT-----GAGAG
Ref. 33_01B.MY.05.05MYKL007_1.DQ366659	TGTGGGAAAAGGAAAGGACATCAAATGAAAAGATTGTACT-----GAGAG
Ref. 33_01B.MY.05.05MYKL045_1.DQ366662	TGTGGGCAGGAAGGACATCAAATGAAAAGATTGTACC-----GAGAG
Ref. 34_01B.TH.99.OUR2478P.EF165541	TGTGGGAAAAGGAAAGGACATCAAATGAAAAGACTGCACT-----GAGAA
Ref. 35_AD.AF.05.05AF026.EF158043	TGTGGGAAAAGAAGGACACCCAAATGAAAAGACTGCACT-----GAGAG
Ref. 35_AD.AF.05.05AF094.EF158040	TGTGGGAAAAGAAGGACACCCAAATGAAAAGACTGCACT-----GAGAG
Ref. 36_cpx.CM.00.00CMNYU1162.EF087995	TGTGGGAAAAGGAAAGGACACCCGAATGAAAAGACTGCACT-----AATGAAAAG
Ref. 36_cpx.CM.00.00CMNYU830.EF087994	TGTGGGAAAAGGAAAGGACACCCAAATGAAAAGACTGCACT-----AATGAAAAG
Ref. 37_cpx.CM.00.00CMNYU926.EF116594	TGTGGAAAAGGAAAGGACACCCAAATGAAAAGACTGCACT-----GAGAG
Ref. 37_cpx.CM.97.CM53392.AF377957	TGTGGGAAAAGGAAAGGACACCCAAATGAAAAGACTGCACT-----GAAAG
Ref. 39_BF.BR.03.03BRRJ103.EU735534	TGTGGAAAAGGAAAGGACACCCAAATGAAAAGAATGCACA-----GAGAG
Ref. 39_BF.BR.03.03BRRJ327.EU735536	TGTGGAAAAGGAAAGGACACCCAAATGAAAAGATTGTGTG-----GAGAG
Ref. 39_BF.BR.04.04BRRJ179.EU735535	TGTGGAAAAGGAAAGGACACCCAAATGCTAGACTGTACT-----GAAAG
Ref. 40_BF.BR.04.04BRRJ115.EU735538	TGTGGAAAGAGAAGGACACCCAAATGAAAAGATTGTACT-----GAGAG
Ref. 40_BF.BR.04.04BRSQ46.EU735540	TGTGGAAAAAGAAGGACACCCAAATGAAAAGATTGTACT-----GAGAG
Ref. 40_BF.BR.05.05BRRJ200.EU735539	TGTGGAAAAAGAAGGACACCCAAATGAAAAGATTGTGAT-----ATGAG
Ref. 42_BF.LU.03.1uBF_05_03.EU170155	TGTGGAAAGAGAGGGACACCCAAATGAAAAGACTGCACT-----GAAAG
Ref. 43_02G.SA.03.J11223.EU697904	TGTGGAAAAGGAGGGACATCAAATGAAAAGACTGCACA-----GAAAG
Ref. 43_02G.SA.03.J11243.EU697907	TGTGGAAAAGGAGGGACATCAAATGAAAAGAATGCACA-----GAGAG
Ref. 43_02G.SA.03.J11456.EU697909	TGTGGAAAAGGAGGGACATCAAATGAAAAGACTGCACA-----GAGAG
Ref. N.CM.02.DJ00131.AY532635	TGTGGCCAAGAAGGACATCAAATGAAAAGATTGAAAAAT-----GAAGGAAG
Ref. N.CM.95.YBF30.AJ006022	TGTGGCCAAGAAGGACATCAAATGAAAAGATTGAAAAAT-----GAAGGAAG
Ref. N.CM.97.YBF106.AJ271370	TGTGGCCAAGAAGGACATCAAATGAAAAGATTGAAAAAT-----GAGGGAAG
Ref. O.BE.87.ANT70.L20587	TGTGGACAGGAAGGTCACCCAAATGAAAAGATTGCAGAAAT-----GAAAA

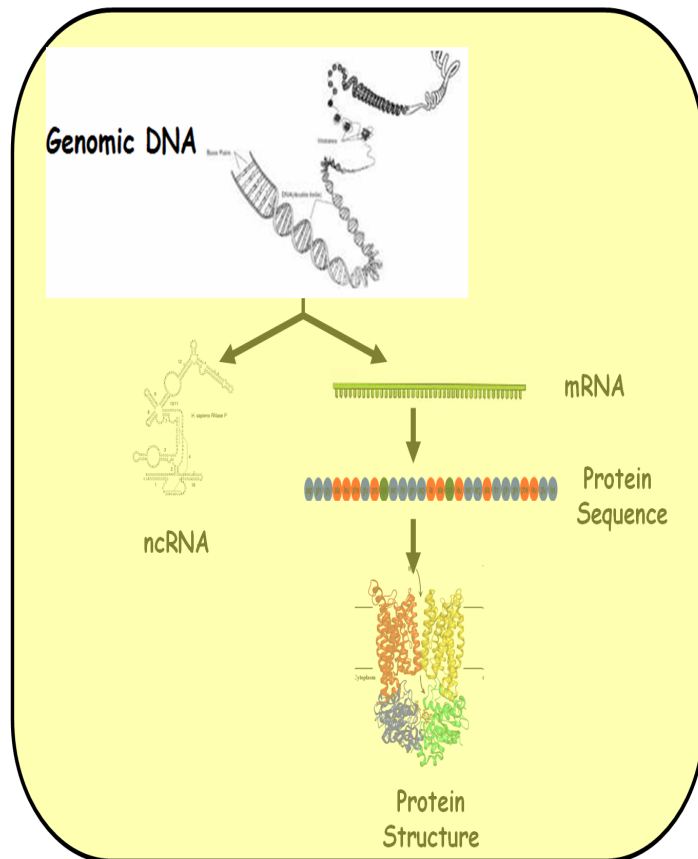
Recent Evolutions

- Consistency
- Model based alignment
- Meta-methods

Which Tool for Which Sequence ?

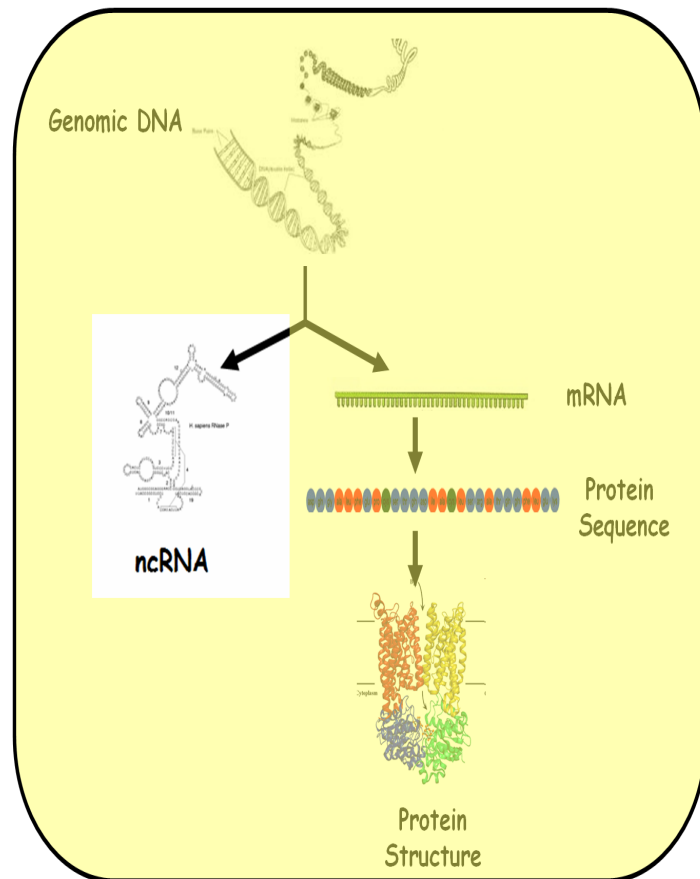


Is it Possible to Compare all Types of Sequences ?



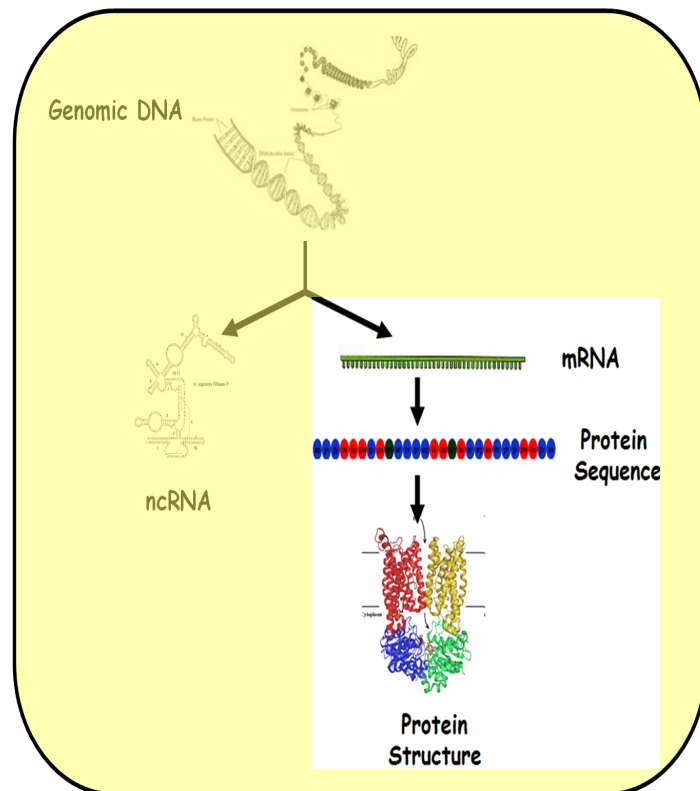
- Non Transcribed World
 - Genes/Full Genomes
 - Lagan, TBA, Pecan
 - Promoter Regions
 - Meta-Aligner
 - Motifs Finders
 - Nucleosome
 - ???
- Multiple Genome Aligners
 - Not Very Accurate
 - Very Fast
 - Deal with rearrangements

Is it Possible to Compare all Types of Sequences ?



- RNA Comparison
 - Less Accurate than Proteins
 - Secondary Structures
- ncRNA World
 - ConSan
 - R-Coffee

Is it Possible to Compare all Types of Sequences ?



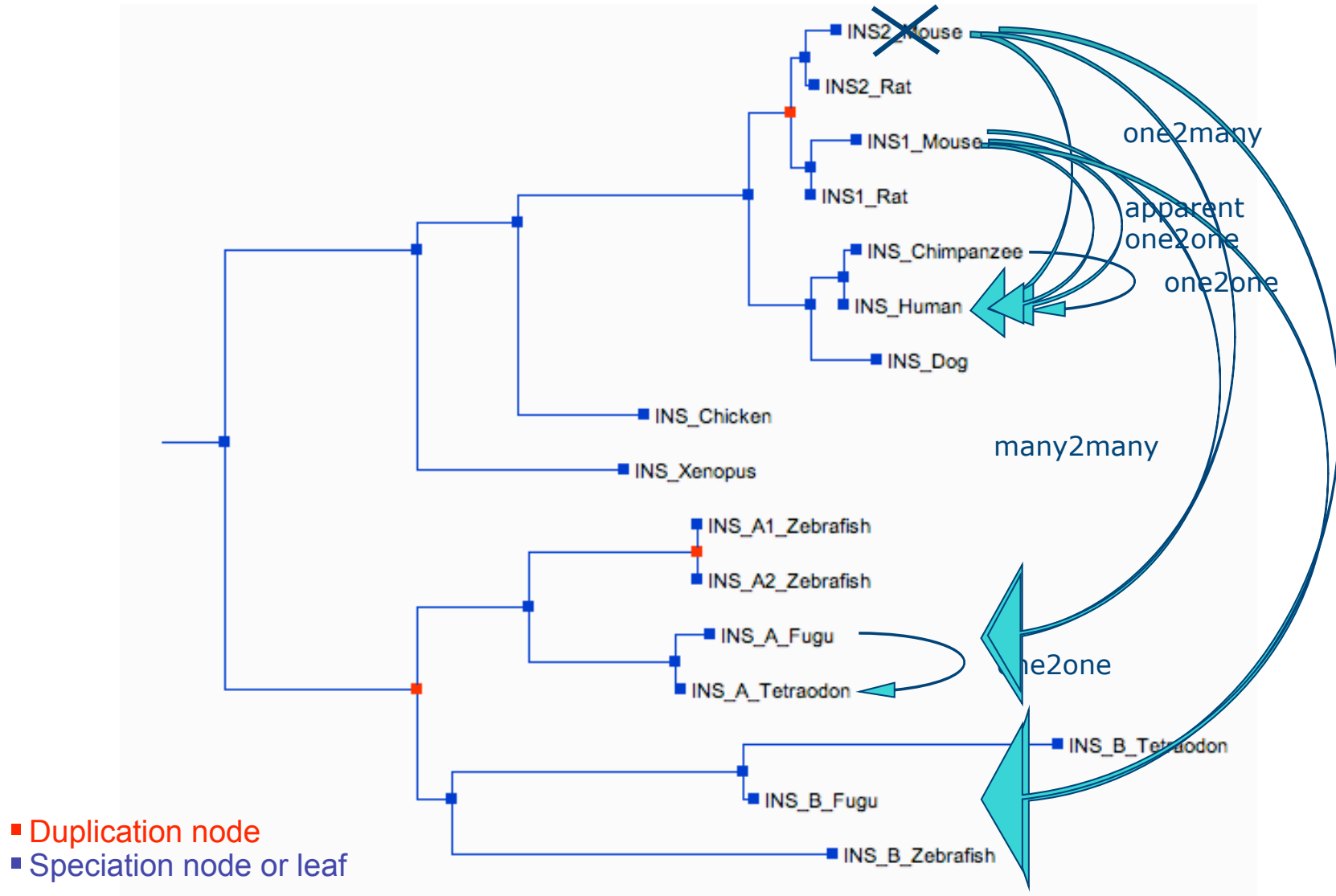
- Protein Comparisons
 - Very Accurate
 - 3D-Structure Improves it
- Protein Aligners
 - ClustalW
 - Muscle
 - Mafft
 - T-Coffee
 - 3D-Coffee

What Changes with 1000 Genomes?

Phylogeny

Phylogeny Vs Function

- Function
 - Low level => Biochemistry => Protein Domains
 - High Level => Metabolic Pathway => Orthology
- Orthology
 - Phylogenetic Analysis
 - Phylogenetic Analysis => Accurate Alignments



(Adpated from “Going beyond AGC and T, E. Birney)

Using The tree

Correct Tree



Correct Orthologous Assignment



Correct Functional Prediction

Trees Vs Alignments

1: [Science](#). 2008 Jan 25;319(5862):473-6.

Alignment uncertainty and genomic analysis.

[Wong KM](#), [Suchard MA](#), [Huelsenbeck JP](#).

1: [Science](#). 2008 Jun 20;320(5883):1632-5.

Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis.

[Löytynoja A](#), [Goldman N](#).

Phylogenetic Trees and Multiple Sequence Alignments

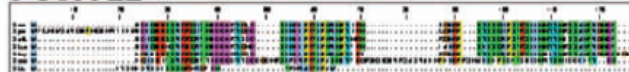
CLUSTAL W



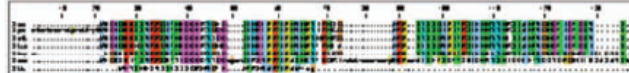
MUSCLE



T-COFFEE



DIALIGN 2



MAFFT



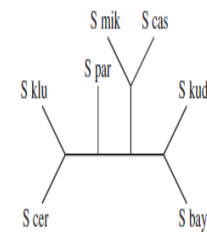
DCA



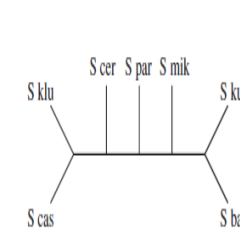
PROBCONS



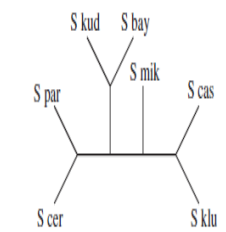
CLUSTAL/DIALIGN (0.24)



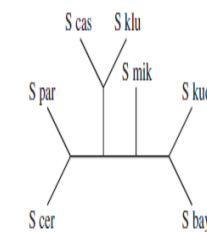
MUSCLE (0.25)



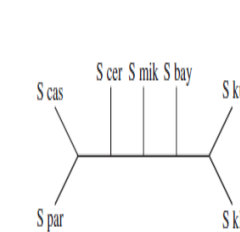
T-COFFEE (0.30)



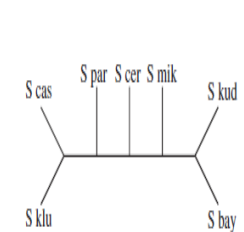
MAFFT (0.18)



DCA (0.12)



PROBCONS (0.05)



Alignment Uncertainty and Genomic Analysis

Genomic Era: The Goal

- 10.000 Sequences: interspecies
- 1 Billion: Re-sequencing
- Incorporation of ALL experimental Data
 - Structure, Genomic, Chlp-Chip, Chlp-Seq...
- Alignments suitable for all applications of comparative genomics
 - Homology Modeling (function)
 - Functional Analysis
 - Phylogenetic Reconstruction
 - 3D-Modelling
- Accurate Alignments for ALL kind of data
 - Non Transcribed DNA
 - Transcribed DNA
 - Translated DNA

Genomic Era Challenges

- Accuracy

- Proteins: 30% is the limit
- DNA/RNA 70% is the limit

- Scale

- Over 100 sequences algorithms lose in accuracy

- Data Integration

- Structure
- Homology
- Genomic Structure
- Function
- Proteomics

- Methods

- Wealth of alternative methods
- Poorly Characterized



Method and Data Integration With Consistency Based Methods

Consistency and Data Integration

- Most methods rely on the progressive algorithm
- Consistency based methods have been designed as an extension
- Consistency based alignment methods have been designed to:
 - Better extract the signal contained in the data
 - Integrate/Confront existing methods
 - Integrate/Confront heterogeneous types of Information

T-Coffee and Concistency...

SeqA GARFIELD THE LAST **FAT** **CAT** Prim. Weight =88

SeqB GARFIELD THE **FAST** **CAT** ---

SeqA GARFIELD THE **LAST** FA-T **CAT** Prim. Weight =77

SeqC GARFIELD THE **VERY** FAST **CAT**

SeqA GARFIELD THE LAST FAT **CAT** Prim. Weight =100

SeqD ----- THE ---- FAT **CAT**

SeqB GARFIELD THE ---- FAST **CAT** Prim. Weight =100

SeqC GARFIELD THE VERY FAST **CAT**

SeqC GARFIELD THE VERY FAST **CAT** Prim. Weight =100

SeqD ----- THE ---- FA-T **CAT**

T-Coffee and Concistency...

```
SeqA GARFIELD THE LAST FAT CAT      Prim. Weight =88
SeqB GARFIELD THE FAST CAT  ---

SeqA GARFIELD THE LAST FA-T CAT      Prim. Weight =77
SeqC GARFIELD THE VERY FAST CAT

SeqA GARFIELD THE LAST FAT CAT      Prim. Weight =100
SeqD ----- THE ---- FAT CAT

SeqB GARFIELD THE ---- FAST CAT      Prim. Weight =100
SeqC GARFIELD THE VERY FAST CAT

SeqC GARFIELD THE VERY FAST CAT      Prim. Weight =100
SeqD ----- THE ---- FA-T CAT
```

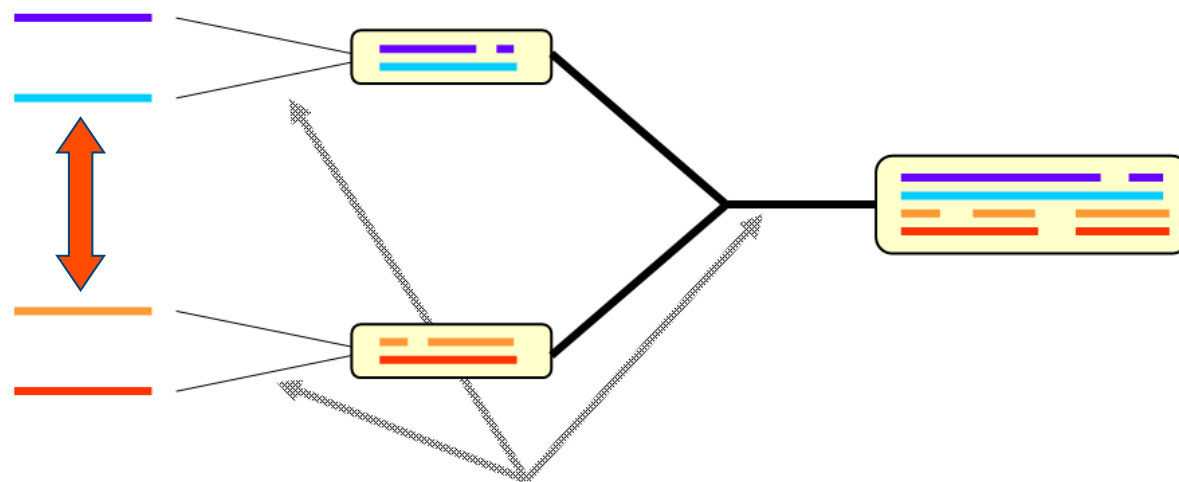
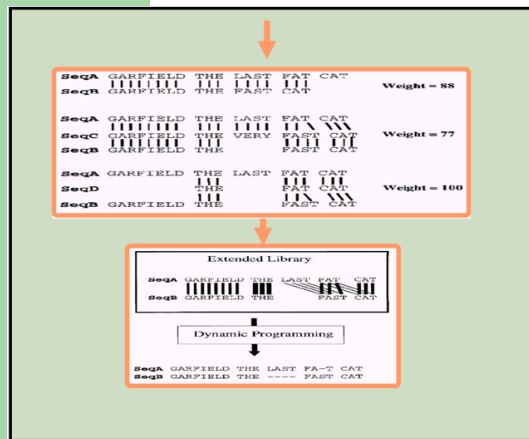


```
SeqA GARFIELD THE LAST FAT CAT      Weight =88
SeqB GARFIELD THE FAST CAT  ---

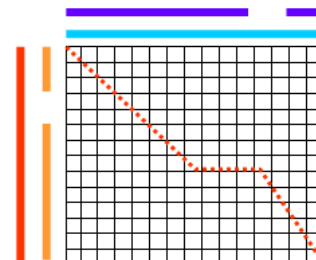
SeqA GARFIELD THE LAST FA-T CAT      Weight =77
SeqC GARFIELD THE VERY FAST CAT
SeqB GARFIELD THE ---- FAST CAT

SeqA GARFIELD THE LAST FA-T CAT      Weight =100
SeqD ----- THE ---- FA-T CAT
SeqB GARFIELD THE ---- FAST CAT
```

T-Coffee and Consistency...

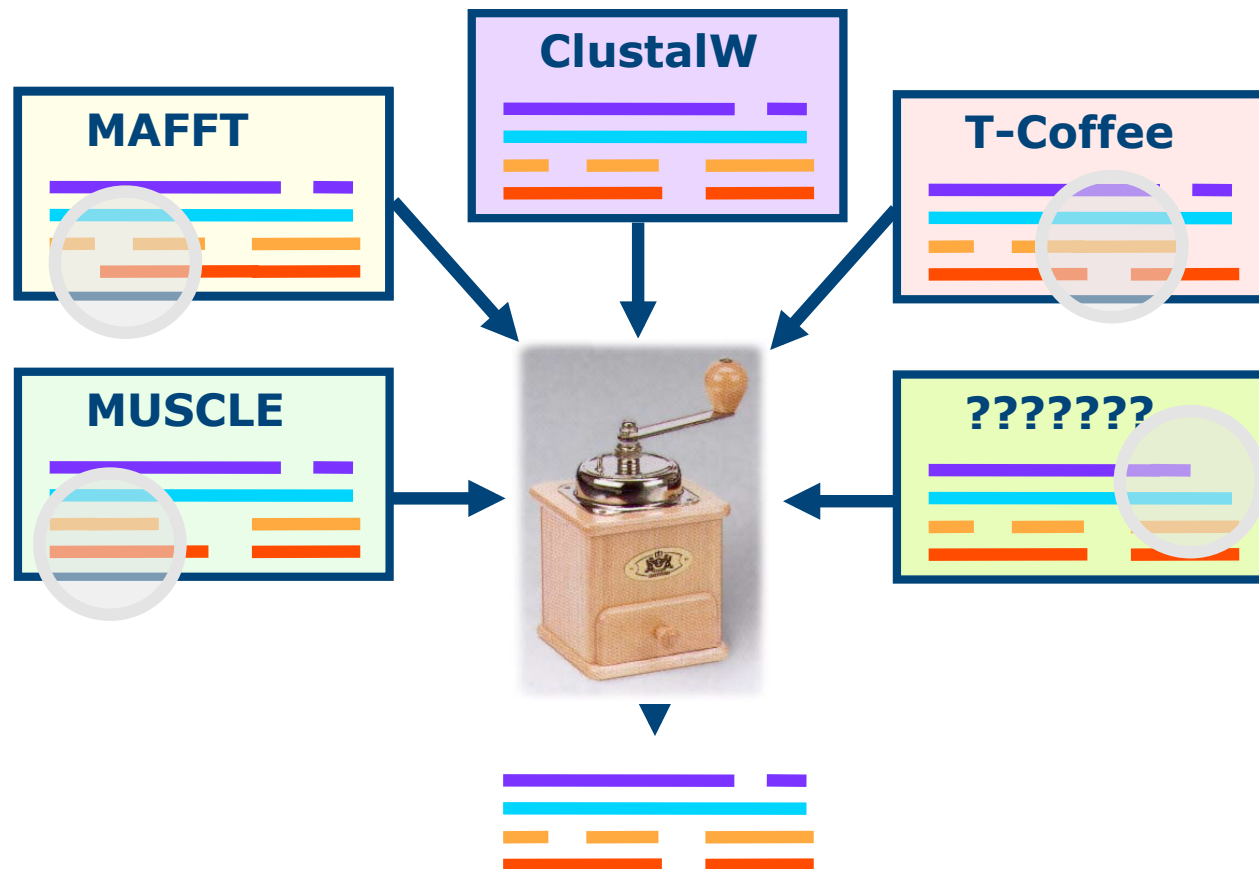


Dynamic Programming Using A Substitution Matrix



M-Coffee

Combining Many MSAs into ONE



Consistency and Accuracy

SCORE=34

*

BAD **AVG** **GOOD**

*

1thx	:	48
lgrx	:	41
lerv	:	47
1a8l	:	28
1ewx A	:	32
1j0f A	:	10
2trc P	:	27
1jfu A	:	38
1kng A	:	35
1sel A	:	28
1mek	:	43
cons	:	34

```

1thx  PCQLMSPLINLAANTYs-----drlkvVKLEIdpn-----
lgrx  YSVRAKDLAEKLSNERd-----dfqyqyvdiraegit-----
lerv  PCKMIKPPFFHSLSEKYS-----nvifLEVVDvdd-----
1a8l  ycplavrmahkfaienkagKqkilqdmveaiey-----
1ewx_A PCRGFTPQLIEFYDKFhesk--nfevVFCTWdeeedgfagyfak-----mpwla
1j0f_A eiksqqsevtrildgkr-----iqyqlvdisqd-----
2trc_P GCDALNSSLECLAAEYpx-----vkfCKIRAsnt-----
1jfu_A PCRKEMPALDELQKLSgp-----nfevVAINIIdtrdpekpktflkeanltrlgyf
1kng_A PCHDEAPLLTELGKdkrf-----qlvginykdaadnarr--flgrygnpfgrvgv
1sel_A ASKEFEKYTFSDpqvqkal--adtvllqanvtandagdv-----
1mek  HCKALAPEYAKAAGKlkaeg--seirlAKVDatee-----

```

cons

```

1thx  ---pttvkky-----KVEGVPALRLVKG-E---QILDSTEGVi-----skdk
lgrx  ---kedlqqkag---KPVETVPQIFvd-----qqhiggyt-----dfaaw
lerv  ---qdvasec-----EVKSMPTFQFFKK-G---QKVGEFSGan-----kek
1a8l  ---pewadqy-----NVMAVPKIIVIQVN-G---EDRVEFEGAY-----pekM
1ewx_A vpfqaseavqklskhFNVESIPTLIGVDA---dsgdvvtt-----ra
1j0f_A ---nalrdemrtlagNPKATPPQIVngn-----hycgd-----yel
2trc_P ---qaqdrf-----SSDVLPTLLVYKG-G---ELISNFISVaEQfAEDffaad
1jfu_A ndqkakovfqdlkaig--RALGMPTSVLVDPQG---CEIATIAGP---aewased
1kng_A dangrasiew-----GVYGVPETFVVGREG--TIVYKLVGP---ITPDnlrsv
1sel_A ---allkhl-----NVLGLPTILFFDGQGEhpqarVTG-----fmdae
1mek  ---sdlagqy-----GVRGYPTIKFFR-nGDTaspkeytag-----readd

```

cons

ENSEMBL-Dev Release 53

Summary of Declaration of intentions

Homologies and families

* Update for the new/updated genebuilds and assembly.

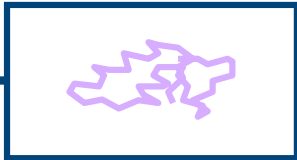
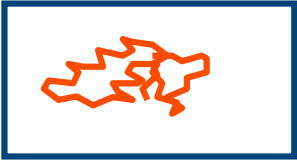
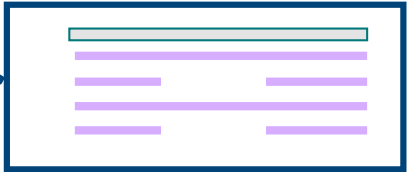
* Replace current clustering method with hcluster for the GeneTrees.

**•Replace current Muscle MSA method with MCoffee.
MCoffee uses a combination of MAFFT-INS, Muscle, Kalign
and Probcons to create a meta-alignment that is a
consensus of all methods.**

* Sitewise dN/dS: we will provide calculation for dN/dS ratios for the(sub)trees that are not dS saturated.

Templates

Templates



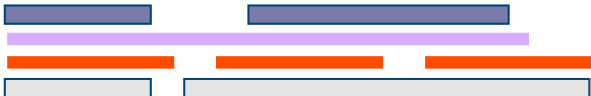
TARGET

TARGET

Template Aligner



Template Alignment



Template-Sequence Alignment



Template based Alignment of the Sequences

Primary Library

Exploring The Template World

Template	Generator	Alignment Method	Mode
RNA Structure	Prediction	RNA Aligner	R-Coffee
Protein Structure	BLAST /PDB	3D Aligner	3D-Coffee
Profile	BLAST/NR	Profile/Profile	PSI-Coffee
Gene Structure	ENSEMBL	Genome Aligner	Exoset
Promoter	Transfac	Meta-Aligner	Meta-Coffee

Method	Method	Template	Score	Comment
ClustalW-2	Progressive	NO	22.74	
PRANK	Gap	NO	26.18	Science2008
MAFFT	Iterative	NO	26.18	
Muscle	Iterative	NO	31.37	
ProbCons	Consistency	NO	40.80	
ProbCons	MonoPhasic	NO	37.53	
T-Coffee	Consistency	NO	42.30	
M-Coffe4	Consistency	NO	43.60	
PSI-Coffee	Consistency	Profile	53.71	
PROMAL	Consistency	Profile	55.08	
PROMAL-3D	Consistency	PDB	57.60	
3D-Coffee	Consistency	PDB	61.00	Espresso

Consistency



Score: fraction of correct columns when compared with a structure based reference (BB11 of BaliBase).

Method	Method	Template	Score	Comment
ClustalW-2	Progressive	NO	22.74	
PRANK	Gap	NO	26.18	Science2008
MAFFT	Iterative	NO	26.18	
Muscle	Iterative	NO	31.37	
ProbCons	Consistency	NO	40.80	
ProbCons	MonoPhasic	NO	37.53	
T-Coffee	Consistency	NO	42.30	
M-Coffe4	Consistency	NO	43.60	
PSI-Coffee	Consistency	Profile	53.71	
PROMAL	Consistency	Profile	55.08	
PROMAL-3D	Consistency	PDB	57.60	
3D-Coffee	Consistency	PDB	61.00	Espresso


Homology Extension



Score: fraction of correct columns when compared with a structure based reference (BB11 of BaliBase).

Method	Method	Template	Score	Comment
ClustalW-2	Progressive	NO	22.74	
PRANK	Gap	NO	26.18	Science2008
MAFFT	Iterative	NO	26.18	
Muscle	Iterative	NO	31.37	
ProbCons	Consistency	NO	40.80	
ProbCons	MonoPhasic	NO	37.53	
T-Coffee	Consistency	NO	42.30	
M-Coffe4	Consistency	NO	43.60	
PSI-Coffee	Consistency	Profile	53.71	
PROMAL	Consistency	Profile	55.08	
PROMAL-3D	Consistency	PDB	57.60	
3D-Coffee	Consistency	PDB	61.00	Espresso

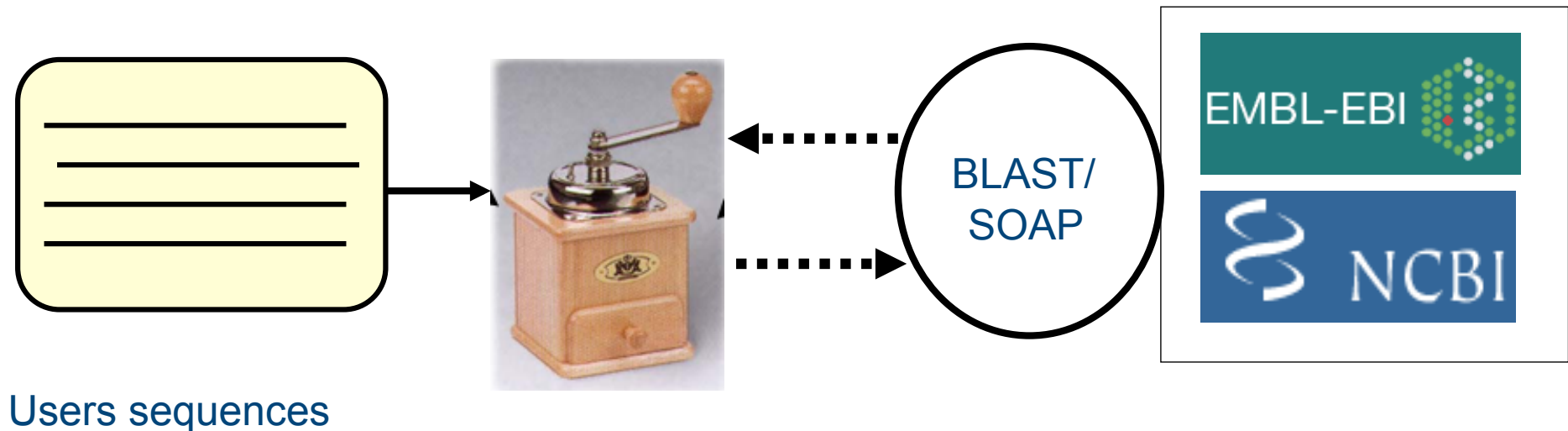
Structural Extension



Score: fraction of correct columns when compared with a structure based reference (BB11 of BaliBase).

T-Coffee and The World

- Some Templates are obtained with a BLAST
- Queries can be sent to the EBI or the NCBI
- No Need for a Local BLAST installation



Genomic Era Challenges

Conclusion

Homology
Extension
(Proteins)

R-Coffee

- Accuracy
 - Proteins: 30% is the limit
 - DNA/RNA 70% is the limit

- Scale
 - Over 100 sequences algorithms lose in accuracy

Scaled Consistency

- Data
 - Structure
 - Homology
 - Genomic Structure
 - Function
 - Proteomics

Template
Based
Alignments

- Methods
 - Wealth of alternative methods
 - Poorly Characterized

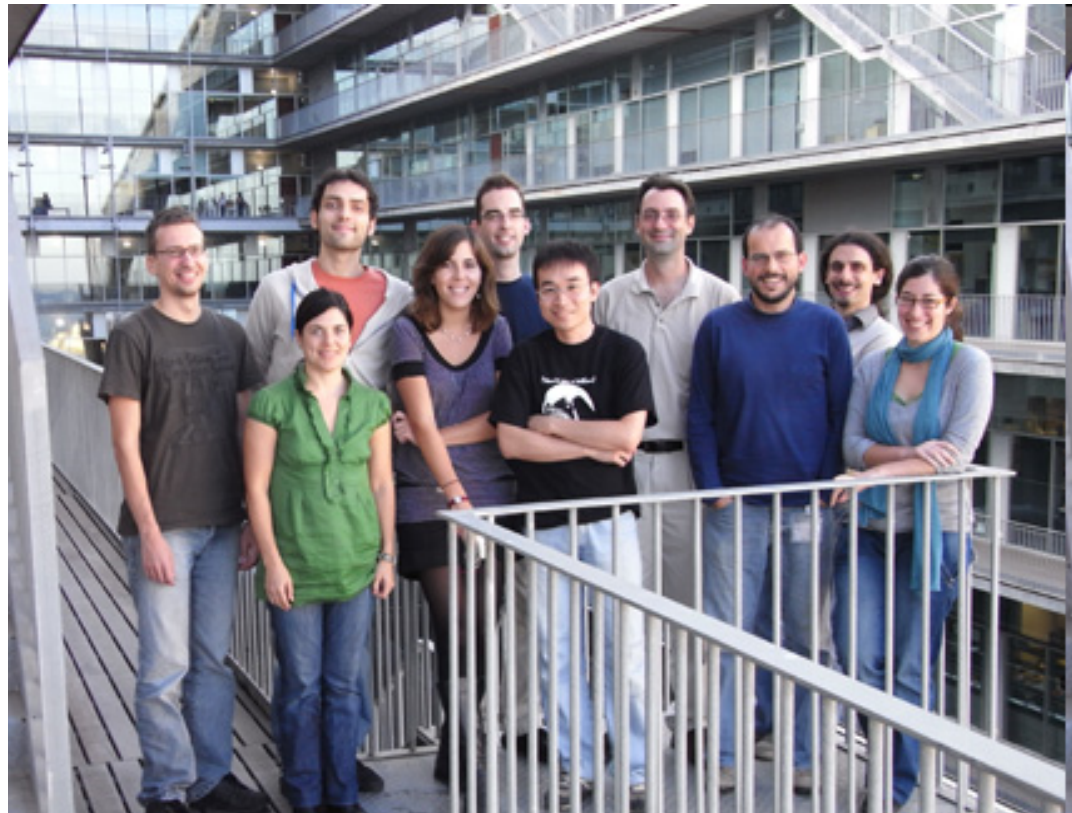
Meta-Methods
M-Coffee

Open Questions

- Accurately Aligning non transcribed DNA
- Accurately aligning ncRNA
- Scaling up consistency based methods with large numbers of sequences
- Coping with Large Number of Re-sequenced Genomes

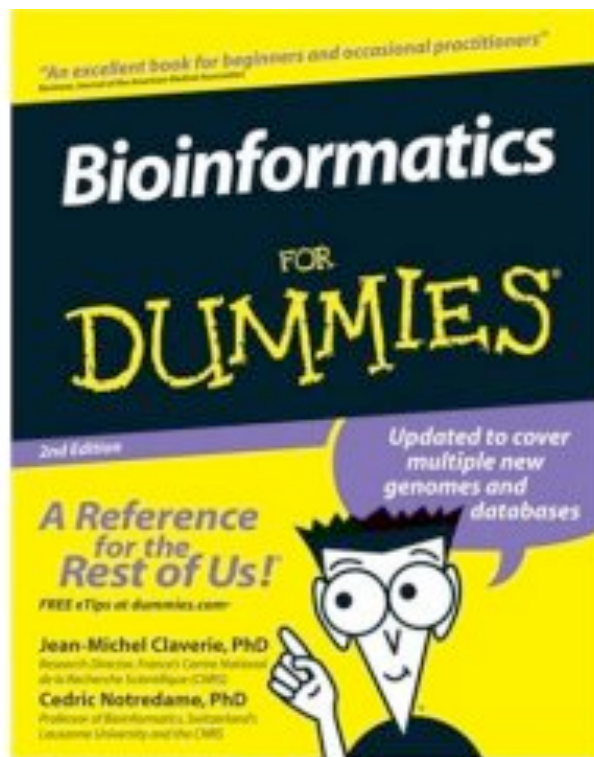
Comparative Bioinformatics

- University College Dublin
 - *Des Higgins*
 - *Orla O'Sullivan*
 - *Iain Wallace (UCD, IE)*
- Berlin Free University
 - Knut Reinert
 - Tobias Rausch
- Swiss Institute of Bioinformatics
 - Ioannis Xenarios
 - Sebastien Morreti
- Comparative Bioinformatics
 - Merixell Oliva
 - Giovanni Bussoti
 - Carsten Kemena
 - Emanuele Rainieri
 - Ionas Erb
 - Jia Ming Chang
 - Matthias Zytneki



www.tcoffee.org
cedric.notredame@crg.es

www.tcoffee.org



Mirror sites:       

ALIGNMENT			
TCOFFEE	<input type="button" value="Regular"/>	<input type="button" value="Advanced"/>	cite ?
EXPRESSO (3DCoFfee)	<input type="button" value="Regular"/>	<input type="button" value="Advanced"/>	cite ?
MCOFFEE	<input type="button" value="Regular"/>	<input type="button" value="Advanced"/>	cite ?
RCOFFEE	<input type="button" value="Regular"/>	<input type="button" value="Advanced"/>	cite ?
COMBINE	<input type="button" value="Regular"/>	<input type="button" value="Advanced"/>	cite ?
EVALUATION			
CORE	<input type="button" value="Regular"/>	<input type="button" value="Advanced"/>	cite ?
iRMSD-APDB	<input type="button" value="Regular"/>	<input type="button" value="Advanced"/>	cite ?
PROCESSING			
PROTOGENE	<input type="button" value="Regular"/>	<input type="button" value="Advanced"/>	cite ?

www.tcoffee.org
cedric.notredame@europe.com

