Global & Local Alignments, Pairwise & Multiple Alignments & Statistical Significance of Alignments

Dr. Abdelkrim Rachedi,

Wits Bioinformatics,
Wits University

Global & Local Alignments

Global Alignment:

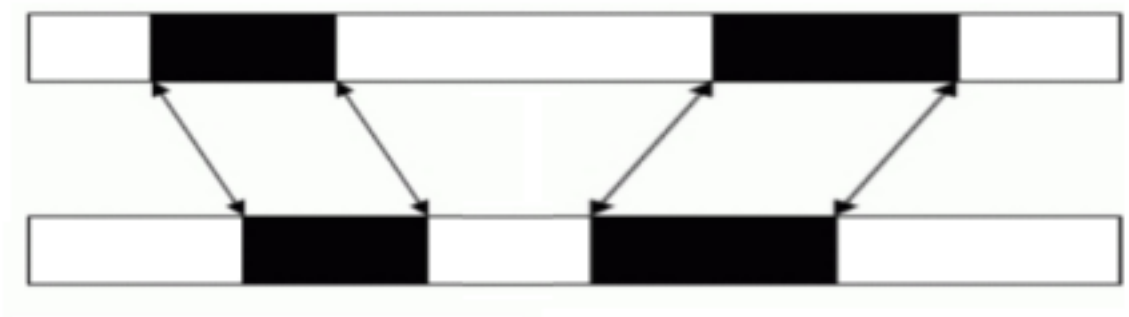
Alignment of the whole sequence 1 with all of the whole sequence 2.

Local Alignment:

Alignment of regions in sequence 1 with regions in sequence 2.

Global & Local Alignments

Sequence Alignment Types



Global

```
FTFTALILLAVAV
F--TAL-LLA-AV
```

Local

```
FTFTALILL-AVAV
--FTAL-LLAAV--
```

Global & Local Alignments

Global alignment (Needleman–Wunsch algorithm, d.prog., 70)

- most useful for “related” sequences and roughly equal length sequences. !!

- Similarity score may not be always good ones.

Global & Local Alignments

Local alignment (Smith-Waterman algorithm, S-W, d.prog, 81) is most used for:

- Difficulties in aligning regions of low similarity.

- Avoids low similarity regions and do those conserved ones. (useful for unrelated seqs. & motifs).

- Accurate & obtains best similarity scores; “optimal alignments”. **However**, it’s time and memory consuming algorithm.

FASTA (an imp. S-W) and BLAST algorithm (a ‘modified S-W’, heuristic) has taken oven for it’s speed (x50) though it does not guaranty finding optimal alignments.

Multiple Sequence Alignment (MSA or MSC)

- MSA is the obvious next step after pairwise alignment.
- Motivation includes:
  - Overall understanding of how different seqs relate to each.
  - Define common features between sequences (**active site residues** .. etc.)
  - Helps in creating Phylogentic Trees
  - Useful for doing structure prediction
- More relevant within a family of proteins.

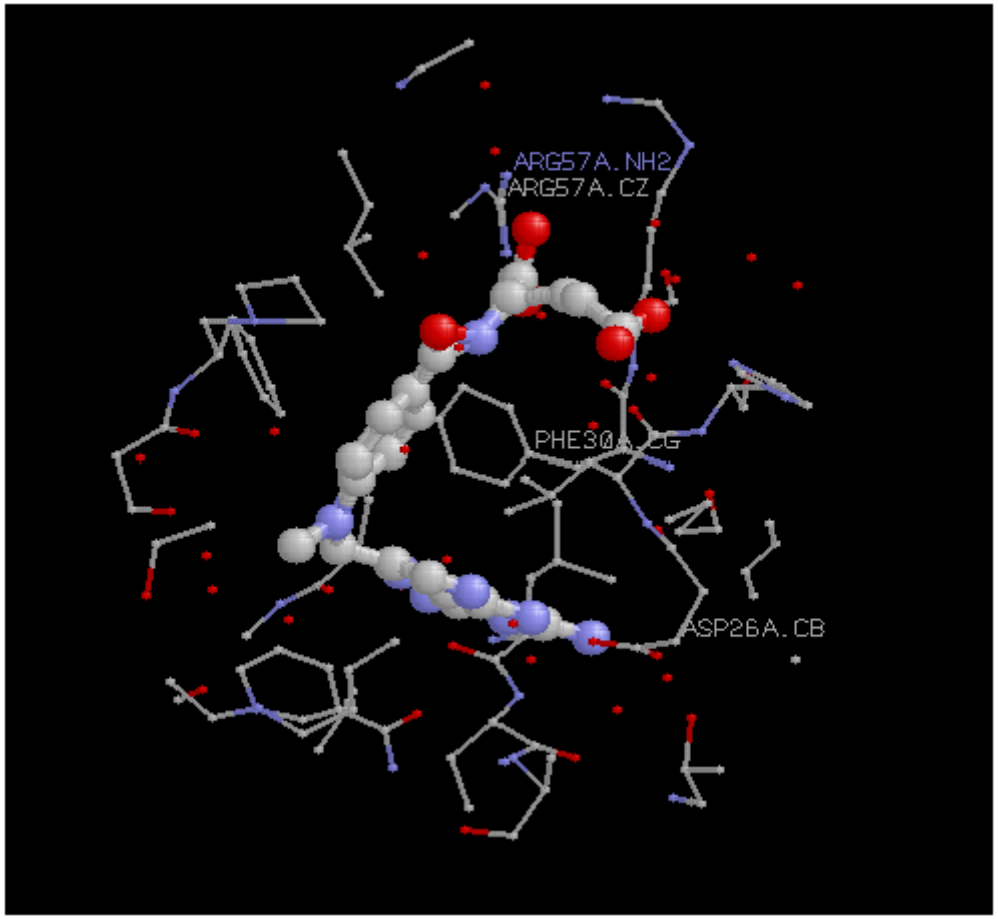
Multiple Sequence Alignment

```
CLUSTAL 2.0.12 multiple sequence alignment


Human           VGSLNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFSRMTTTSSVEGKQNLVIMGKKTWFSI 60
Mouse           VRPLNCIVAVSQNMGIGKNGDLPWPPLRNEFKYFQRMTTTSSVEGKQNLVIMGRKTWFSI 60
Chicken         VRSLNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQNAVIMGKKTWFSI 60
L.Casei         ---TAFLWAQNRNGLIGKDGHLPWH-LPDDLHYFRAQTVG-------KIMVVGRRTYESF 49
                      : *  :*  ***:*.***  * :: :**   *         : :::*::*: *:

Human           PEKSRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSSV 120
Mouse           PEKNRPLKDRINIVLSRELKEPPRGAHFLAKSLDDALRLIEQPELASKVDMVWIVGGSSV 120
Chicken         PEKNRPLKDRINIVLSRELKEAPKGAHYLSKSLDDALALLDSPELKSKVDMVWIVGGTAV 120
L.Casei         PK--RPLPERTNVVLTHQEDYQAQGAVVVHD-----VAAVFAYAKQHLDQELVIAGGAQI 102
                *:  ***  * *:**::: .  .:**  :       :            : : *.**: :

Human           YKEAMNHPGHLKLFVTRIMQDFESDTFFPEIDLEKYKLLPEYPGVLSDVQEEKGIKYKFE 180
Mouse           YQEAMNQPGHLRLFVTRIMQEFESDTFFPEIDLGKYKLLPEYPGVLSEVQEEKGIKYKFE 180
Chicken         YKAAMEKPINHRLFVTRILHEFESDTFFPEIDYKDFKLLTEYPGVPADIQEEDGIQYKFE 180
L.Casei         FTAFKDD--VDTLLVTRLAGSFEGDTKMIPLNWDDFTKVSSR----TVEDTNPALTHTYE 156
                :    :.     *:***:  .**.** :  ::  .:. :..     :  : : .: :.:*

Human           VYEKND--- 186
Mouse           VYEKKD--- 186
Chicken         VYQKSVLAQ 189
L.Casei         VWQKKA--- 162
                *::*.
```

MSA of DHFR sequences

ClustalW2 output (EBI tools)

e.g. Active Site query



L.Casei MTX Binding site

e.g. Active Site query

```
CLUSTAL 2.0.12 multiple sequence alignment


Human           VGSLNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFSRMTTTSSVEGKQNLVIMGKKTWFSI 60
Mouse           VRPLNCIVAVSQNMGIGKNGDLPWPPLRNEFKYFQRMTTTSSVEGKQNLVIMGRKTWFSI 60
Chicken         VRSLNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQNAVIMGKKTWFSI 60
L.Casei         ---TAFLWAQNRNGLIGKDGHLPWH-LPDDLHYFRAQTVG-------KIMVVGRRTYESF 49
                      : *  :*  ***:*.***  * :: :**   *         : :::*::*: *:

Human           PEKSRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSSV 120
Mouse           PEKNRPLKDRINIVLSRELKEPPRGAHFLAKSLDDALRLIEQPELASKVDMVWIVGGSSV 120
Chicken         PEKNRPLKDRINIVLSRELKEAPKGAHYLSKSLDDALALLDSPELKSKVDMVWIVGGTAV 120
L.Casei         PK--RPLPERTNVVLTHQEDYQAQGAVVVHD-----VAAVFAYAKQHLDQELVIAGGAQI 102
                *:  ***  * *:**::: .  .:**  :       :            : : *.**: :

Human           YKEAMNHPGHLKLFVTRIMQDFESDTFFPEIDLEKYKLLPEYPGVLSDVQEEKGIKYKFE 180
Mouse           YQEAMNQPGHLRLFVTRIMQEFESDTFFPEIDLGKYKLLPEYPGVLSEVQEEKGIKYKFE 180
Chicken         YKAAMEKPINHRLFVTRILHEFESDTFFPEIDYKDFKLLTEYPGVPADIQEEDGIQYKFE 180
L.Casei         FTAFKDD--VDTLLVTRLAGSFEGDTKMIPLNWDDFTKVSSR----TVEDTNPALTHTYE 156
                :    :.     *:***:  .**.** :  ::  .:. :..     :  : : .: :.:*

Human           VYEKND--- 186
Mouse           VYEKKD--- 186
Chicken         VYQKSVLAQ 189
L.Casei         VWQKKA--- 162
                *::*.
```

Multiple Sequence Alignment

Phylogram

ClustalW2 output (EBI tools)

Statistical Significance of Alignments

- Percent measure (>= 25% similarity → relatedness between two sequences.)

- Alignment Score

  BLAST use: raw & bit scores

  - Raw is a score value based on the used scoring matrix.
  - Bit score is a normalized value with respect to the scoring system. Bit score is comparable for searches using different scoring matrices.

- Expectancy value or E-value. (< Statistical Significance) – dbase size reliant.

- Propability value or P-Value (seq. To appear by chance)

Statistical Significance of Alignments

```
>UNIPROT:Q78EH1_9RODE Q78EH1 Dihydrofolate reductase OS=Cricetulus sp. PE=2
            SV=1
        Length = 187

 Score = 875 (313.1 bits), Expect = 4.1e-86, P = 4.1e-86
 Identities = 165/184 (89%), Positives = 176/184 (95%)

Query:     1 VGSLNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFSRMTTTSSVEGKQNLVIMGKKTWFSI 60
             V  LNCIVAVSQNMGIGKNGD PWP LRNEF+YF RMTTTSSVEGKQNLVIMG+KTWFSI
Sbjct:     2 VRPLNCIVAVSQNMGIGKNGDFPWPMLRNEFKYFQRMTTTSSVEGKQNLVIMGRKTWFSI 61

Query:    61 PEKSRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSSV 120
             PEK+RPLK RIN+VLSRELKEPPQGAHFL++SLD+ALKL EQPELA+KVDMVWIVGGSSV
Sbjct:    62 PEKNRPLKDRINIVLSRELKEPPQGAHFLAKSLDNALKLIEQPELADKVDMVWIVGGSSV 121

Query:   121 YKEAMNHPGHLKLFVTRIMQDFESDTFFPEIDLEKYKLLPEYPGVLSDVQEEKGIKYKFE 180
             YKEAMN PGHL+LFVTRIMQ+FESDTFFPEIDLEKYKLLPEYPGVLS+VQEEKGIKYKFE
Sbjct:   122 YKEAMNQPGHLRLFVTRIMQEFESDTFFPEIDLEKYKLLPEYPGVLSEVQEEKGIKYKFE 181

Query:   181 VYEK 184
             VYEK
Sbjct:   182 VYEK 185
```

e.g. of BLAST results (run from EBI tools)

Multi-Structural Sequence Alignment

```
Structure  0       3F8Z:A:          1    186
Structure  1       3D80:A:          1    186
Structure  2       8DFR:A:          2    186
Structure  3       3DFR:A:          1    160


Alignment after MC optimization
Final Alignment Distance:    1.005
Final Alignment Score:  3242.20
Final Alignment length:  186 & Total length:  188

  0 3F8Z:A      0 VGSLNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFSRMTTTSSVEGKQNLVIMGKKTWFSIPEKSRPLKGR
  1 3D80:A      0 VRPLNCIVAVSQNMGIGKNGDLPWPPLRNEFKYFQRMTTTSSVEGKQNLVIMGRKTWFSIPEKNRPLKDR
  2 8DFR:A      0 -RSLNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQNAVIMGKKTWFSIPEKNRPLKDR
  3 3DFR:A      0 ---TAFLWAQNRNGLIGKDGHLPW-HLPDDLHYFRAQTV-------GKIMVVGRRTYESFP--KRPLPER

  0 3F8Z:A     70 INLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSSVYKEAMNHPGHLKLFVTRIMQ
  1 3D80:A     70 INIVLSRELKEPPRGAHFLAKSLDDALRLIEQPELASKVDMVWIVGGSSVYQEAMNQPGHLRLFVTRIMQ
  2 8DFR:A     70 INIVLSRELKEAPKGAHYLSKSLDDALALLDSPELKSKVDMVWIVGGTAVYKAAMEKPINHRLFVTRILH
  3 3DFR:A     70 TNVVLTHQEDYQAQGA-VVVHDVAAVFAYAKQ--HL--DQELVIAGGAQIFTAFKDD--VDTLLVTRLAG

  0 3F8Z:A    140 DFESDTFFPEIDLEKYKLLPEYPGVLSDVQEEKG--IKYKFEVYEKND
  1 3D80:A    140 EFESDTFFPEIDLGKYKLLPEYPGVLSEVQEEKG--IKYKFEVYEKKD
  2 8DFR:A    140 EFESDTFFPEIDYKDFKLLTEYPGVPADIQEEDG--IQYKFEVYQKSV
  3 3DFR:A    140 SFEGDTKMIPLNWDDFTKVS------SRTVEDTNPALTHTYEVWQK--
```

Multi-Structural Sequence Alignment

Performed by CEMC - Multiple Protein Structure Alignment Server. http://pathway.rit.albany.edu/~cemc/

Multiple Sequence Alignment

```
CLUSTAL 2.0.12 multiple sequence alignment


Human           VGSLNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFSRMTTTSSVEGKQNLVIMGKKTWFSI 60
Mouse           VRPLNCIVAVSQNMGIGKNGDLPWPPLRNEFKYFQRMTTTSSVEGKQNLVIMGRKTWFSI 60
Chicken         VRSLNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQNAVIMGKKTWFSI 60
L.Casei         ---TAFLWAQNRNGLIGKDGHLPWH-LPDDLHYFRAQTVG-------KIMVVGRRTYESF 49
                      : *  :*  ***:*.***  * :: :**   *         : :::*::*: *:

Human           PEKSRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSSV 120
Mouse           PEKNRPLKDRINIVLSRELKEPPRGAHFLAKSLDDALRLIEQPELASKVDMVWIVGGSSV 120
Chicken         PEKNRPLKDRINIVLSRELKEAPKGAHYLSKSLDDALALLDSPELKSKVDMVWIVGGTAV 120
L.Casei         PK--RPLPERTNVVLTHQEDYQAQGAVVVHD-----VAAVFAYAKQHLDQELVIAGGAQI 102
                *:  ***  * *:**::: .  .:**  :       :            : : *.**: :

Human           YKEAMNHPGHLKLFVTRIMQDFESDTFFPEIDLEKYKLLPEYPGVLSDVQEEKGIKYKFE 180
Mouse           YQEAMNQPGHLRLFVTRIMQEFESDTFFPEIDLGKYKLLPEYPGVLSEVQEEKGIKYKFE 180
Chicken         YKAAMEKPINHRLFVTRILHEFESDTFFPEIDYKDFKLLTEYPGVPADIQEEDGIQYKFE 180
L.Casei         FTAFKDD--VDTLLVTRLAGSFEGDTKMIPLNWDDFTKVSSR----TVEDTNPALTHTYE 156
                :    :.     *:***:  .**.** :  ::  .:. :..     :  : : .: :.:*

Human           VYEKND--- 186
Mouse           VYEKKD--- 186
Chicken         VYQKSVLAQ 189
L.Casei         VWQKKA--- 162
                *::*.
```

MSA of DHFR sequences

ClustalW2 output (EBI tools)

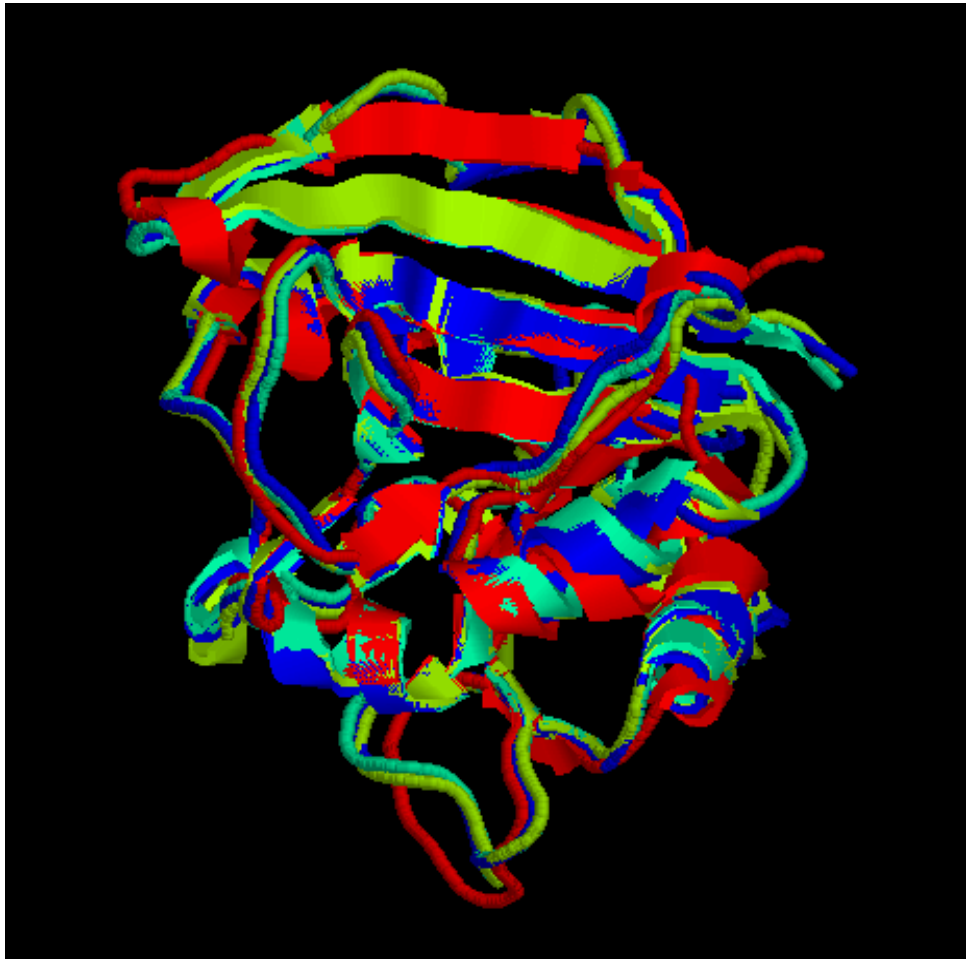
Multi-Structural Sequence Alignment

```
Structure  0       3F8Z:A:          1    186
Structure  1       3D80:A:          1    186
Structure  2       8DFR:A:          2    186
Structure  3       3DFR:A:          1    160


Alignment after MC optimization
Final Alignment Distance:    1.005
Final Alignment Score:  3242.20
Final Alignment length:  186 & Total length:  188

  0 3F8Z:A      0 VGSLNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFSRMTTTSSVEGKQNLVIMGKKTWFSIPEKSRPLKGR
  1 3D80:A      0 VRPLNCIVAVSQNMGIGKNGDLPWPPLRNEFKYFQRMTTTSSVEGKQNLVIMGRKTWFSIPEKNRPLKDR
  2 8DFR:A      0 -RSLNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQNAVIMGKKTWFSIPEKNRPLKDR
  3 3DFR:A      0 ---TAFLWAQNRNGLIGKDGHLPW-HLPDDLHYFRAQTV-------GKIMVVGRRTYESFP--KRPLPER

  0 3F8Z:A     70 INLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSSVYKEAMNHPGHLKLFVTRIMQ
  1 3D80:A     70 INIVLSRELKEPPRGAHFLAKSLDDALRLIEQPELASKVDMVWIVGGSSVYQEAMNQPGHLRLFVTRIMQ
  2 8DFR:A     70 INIVLSRELKEAPKGAHYLSKSLDDALALLDSPELKSKVDMVWIVGGTAVYKAAMEKPINHRLFVTRILH
  3 3DFR:A     70 TNVVLTHQEDYQAQGA-VVVHDVAAVFAYAKQ--HL--DQELVIAGGAQIFTAFKDD--VDTLLVTRLAG

  0 3F8Z:A    140 DFESDTFFPEIDLEKYKLLPEYPGVLSDVQEEKG--IKYKFEVYEKND
  1 3D80:A    140 EFESDTFFPEIDLGKYKLLPEYPGVLSEVQEEKG--IKYKFEVYEKKD
  2 8DFR:A    140 EFESDTFFPEIDYKDFKLLTEYPGVPADIQEEDG--IQYKFEVYQKSV
  3 3DFR:A    140 SFEGDTKMIPLNWDDFTKVS------SRTVEDTNPALTHTYEVWQK--
```

Evaluation of Structural Alignment

```
Structure  0      3F8Z:A:        1    186
Structure  1      3D80:A:        1    186
Structure  2      8DFR:A:        2    186
Structure  3      3DFR:A:        1    160

Alignment after MC optimization
Final Alignment Distance:    1.005
Final Alignment Score:  3242.20
Final Alignment length:  186 & Total length:  188
```

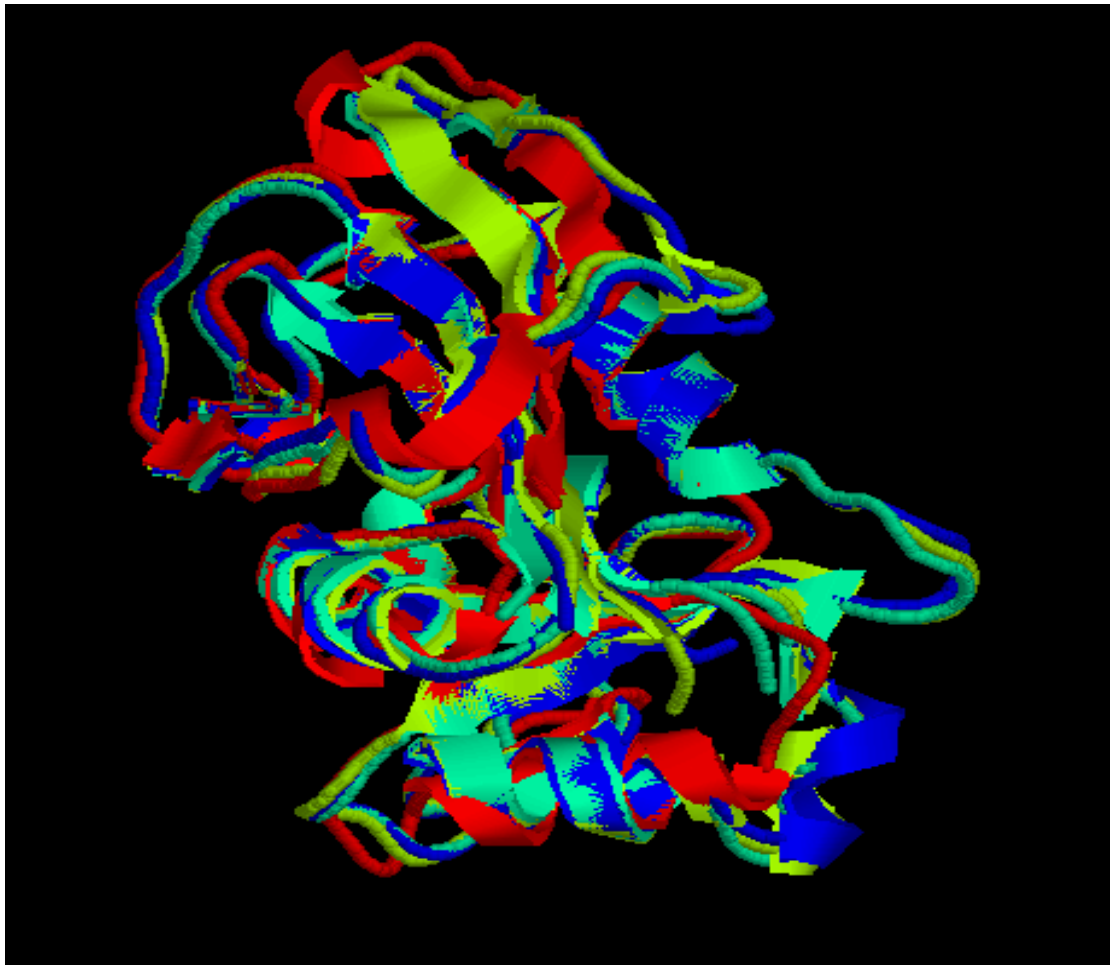
Multi-Structural Alignment



Human, Mouse, Chicken & L.Casei

DHFR Structures fitted with RMSD=1.005 A

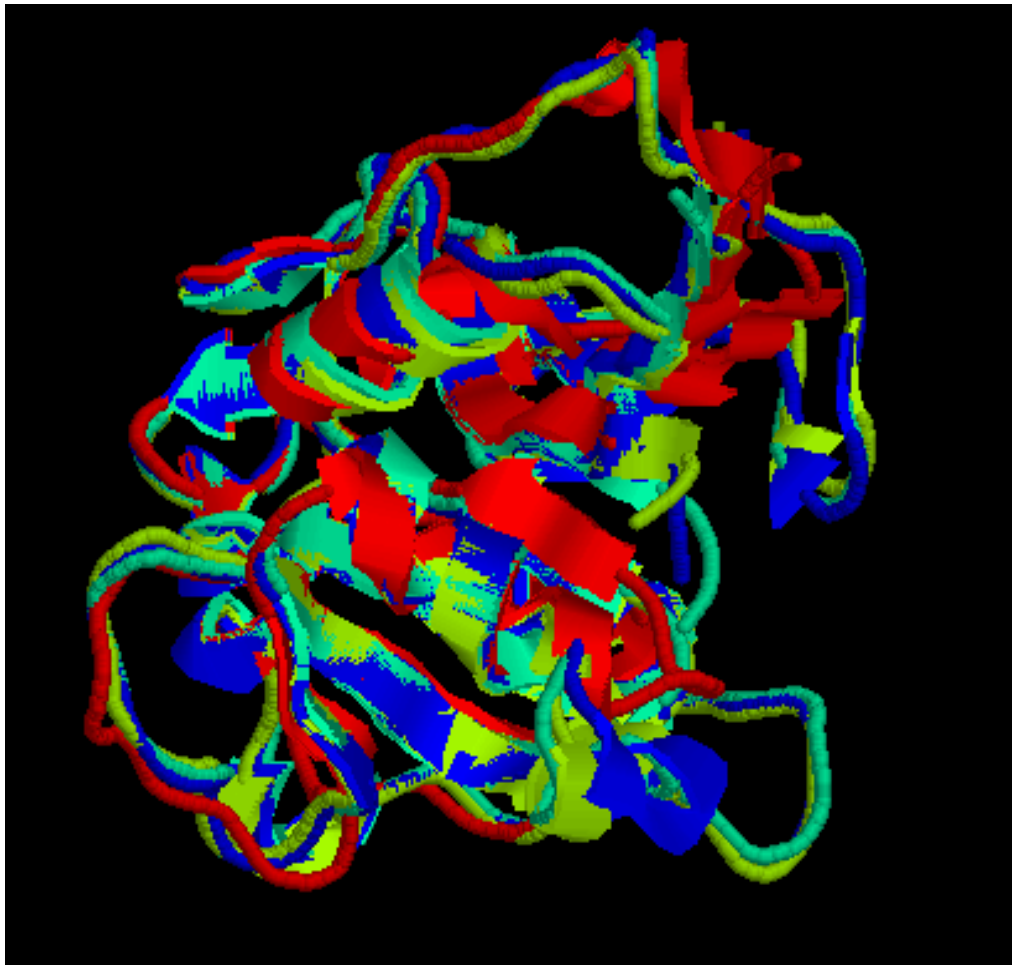
Multi-Structural Alignment



Human, Mouse, Chicken & L.Casei

DHFR Structures fitted with RMSD=1.005 A

Multi-Structural Alignment



Human, Mouse, Chicken & L.Casei

DHFR Structures fitted with RMSD=1.005 A

Alignment Tools

BLAST

Align

ClustalW

MAFFT

MUSCLE

T-Coffee Aligns and combines too.

Align

This tool is used to compare 2 sequences (pairwise).

Global :When you want an alignment that covers the whole length of both sequences, use **needle (Needleman-Wunsch)**.

Local: When you are trying to find the best region of similarity between two sequences, use **water. (Smith-Waterman)**

MUSCLE stands for MUltiple Sequence Comparison by Log-Expectation. MUSCLE is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee, depending on the chosen options.

ClustalW2 is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms.

T-Coffee

T-Coffee is a multiple sequence alignment program. Multiple sequence alignment programs are meant to align a set of sequences previously gathered using other programs such as blast, fast, sw ...

The main characteristic of T-Coffee is that it will allow you to combine results obtained with several alignment methods. For instance if you have an alignment coming from ClustalW2, an other alignment coming from Dialign, and a structural alignment of some of your sequences, T-Coffee will combine all that information and produce a new multiple sequence having the best agreement whith all these methods.

By default, T-Coffee will compare all you sequences two by two, producing a global alignment and a series of local alignments (using lalign). The program will then combine all these alignments into a multiple alignment.

MAFFT (Multiple Alignment using Fast Fourier Transform) is a high speed multiple sequence alignment program.