DIALIGN

A segment-segment multiple sequences alignment program

Kai Ye & Qilan Li

Outline of method section

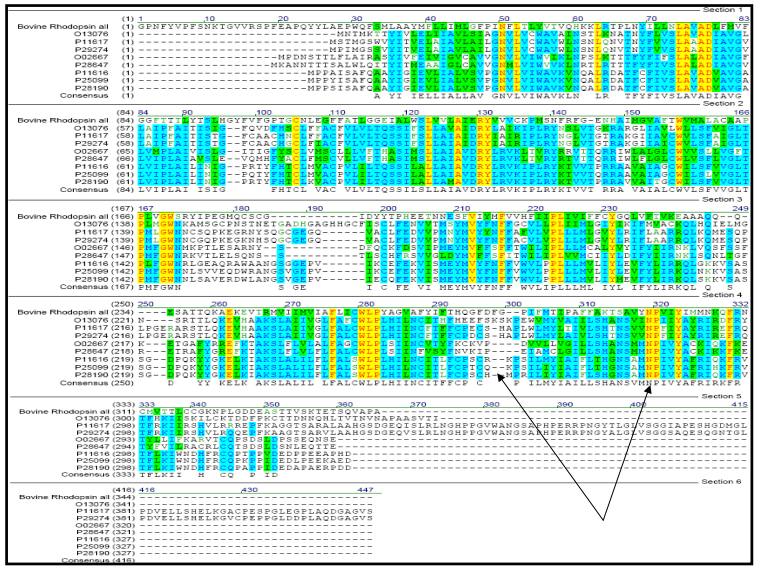
- 1. What is protein sequence?
- 2. What is sequence alignment?
- 3. Why we need sequence alignment?
- 4. Methods to do sequence alignment.
- 5. Advantage and disadvantage of traditional method: ClustalW
- 6. DIALIGN
- 7. Advantage and disadvantage of DIALIGN
- 8. Improvement of speed

What is protein sequences?

- A protein sequence is represented by a string a of letters coding for the 20 different types of amino acid residues.
- >P04201
- MDGSNVTSFVVEEPTNISTGRNASVGNAHRQIPIVHW
  VIMSISPVGFVENGILLWFLCFRMRRNPFTVYITHLSIA
  DISLLFCIFILSIDYALDYELSSGHYYTIVTLSVTFLFGYN
  TGLYLLTAISVERCLSVLYPIWYRCHRPKYQSALVCAL
  LWALSCLVTTMEYVMCIDREEESHSRNDCRAVIIFIAIL
  SFLVFTPLMLVSSTILVVKIRKNTWASHSSKLYIVIMVTII
  IFLIFAMPMRLLYLLYYEYWSTFGNLHHISLLFSTINSSA
  NPFIYFFVGSSKKKRFKESLKVVLTRAFKDEMQPRRQ
  KDNCNTVTVETVV

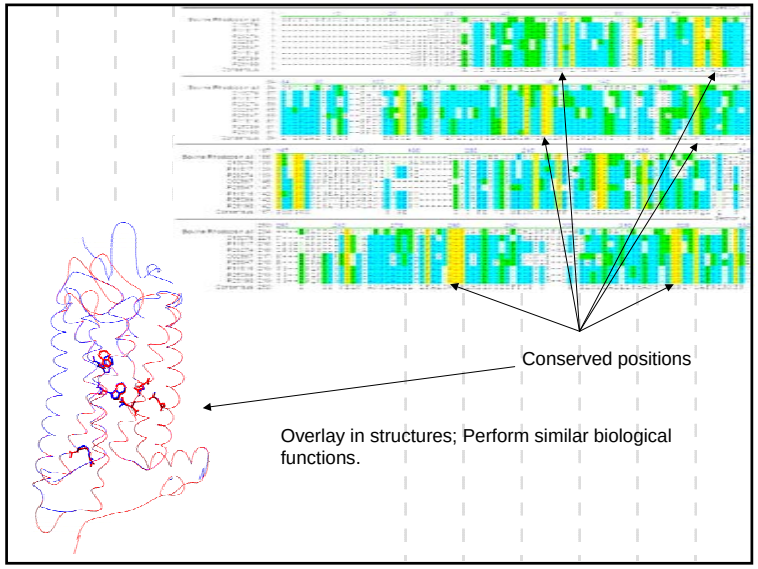
What is sequence alignment?

- A protein sequence alignment is created when the residues in one sequence are lined up with those in at least one other sequence.
- Optimal alignment of the two sequences will usually require the insertion of gaps in one or both sequences in order to find the best alignment.



Why we need sequence alignment?

- Alignment of two residues implies that those residues are performing similar roles in the two different proteins.
- This allows for information known about specific residues in one sequence to be potentially transferred to the residues aligned in the other.



Methods to make sequence alignment

- Global progressive alignment algorithm:
  MULTALIGN; MULTAL; PILEUP; CLUSTALW
  differ mainly in the method used to determine the order of alignment of the sequences.
- Local dynamic programming algorithm
  PIMA;

Traditional methods

Methods to make sequence alignment

- Segment-segment comparisons rather than residue-to-residue comparisons: DIALIGN
- Genetic algorithm: SAGA
- Hidden Markov models (HMMs): HMMT

Speech recognition

"New" methods

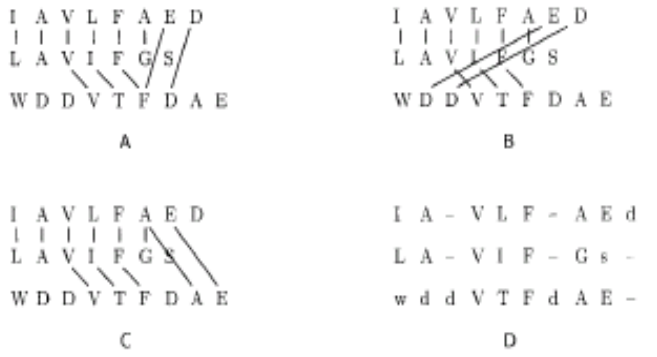
Advantage and disadvantage of traditional method: ClustalW

- ClustalW is superior on globally related protein families (same ancestor).
- However, many protein families share only isolated regions of local similarity.
- The resulting alignments depend sensitively on a set of user-defined parameters, especially the gap penalty.

DIALIGN

- Segment-to-segment alignments with variable segment length
- No gap penalties
- Regions of low similarity are excluded from DIALIGN alignment.

DIALIGN
consistent collections of segment pairs



Advantage and disadvantage of DIALIGN

- Alignments produced by DIALIGN seem to be comparable to the results of standard global alignment methods as CLUSTALW. However, if sequences are only locally related, DIALIGN seems to be clearly superior to other methods.
- Too slow to align sequences in the order of hundreds of kilobases or more.

Improvement of speed

- Solution 1:

  CHAOS, a novel algorithm for rapid identification of chains of local pair-wise sequence similarities. Alignments calculated by CHAOS are used as anchor points to improve the running time of DIALIGN. 95%

Improvement of speed

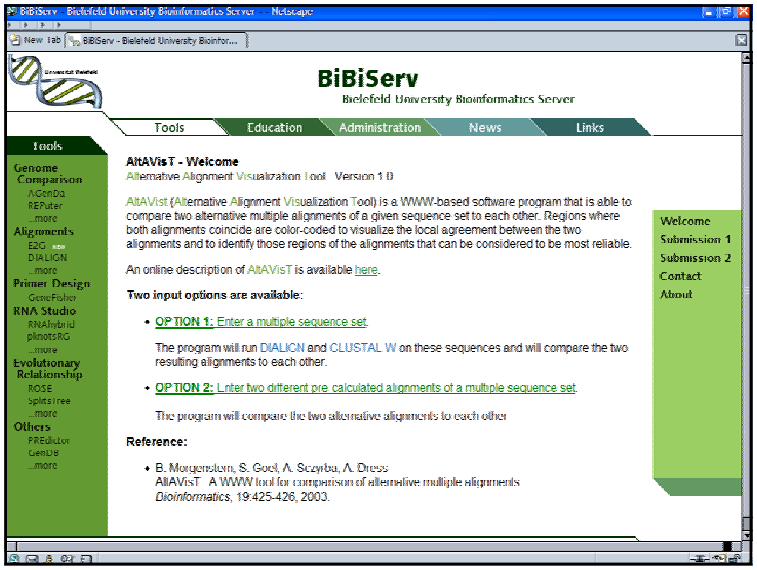
- **Solution 2: DIALIGN P, a parallel version of the multi-alignment program DIALIGN.**
- **A lot of sequences**
- (*a*) pair-wise sequence alignments that are used as a first step to multiple alignment account for most of the CPU time in DIALIGN. Since alignments of different sequence pairs are completely independent of each other, they can be distributed to multiple processors without any effect on the resulting output alignments.
- **Longe sequences**
- (*b*) For alignments of large genomic sequences, DIALIGN use a heuristics by splitting up sequences into sub-sequences based on a previously introduced *anchored alignment* procedure. For test sequences, this combined approach reduces the program running time of DIALIGN by up to 97%.

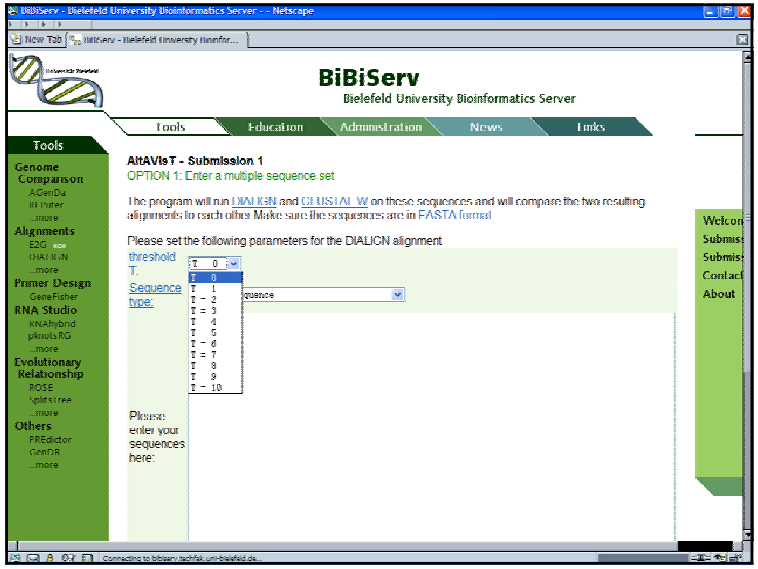
Web server

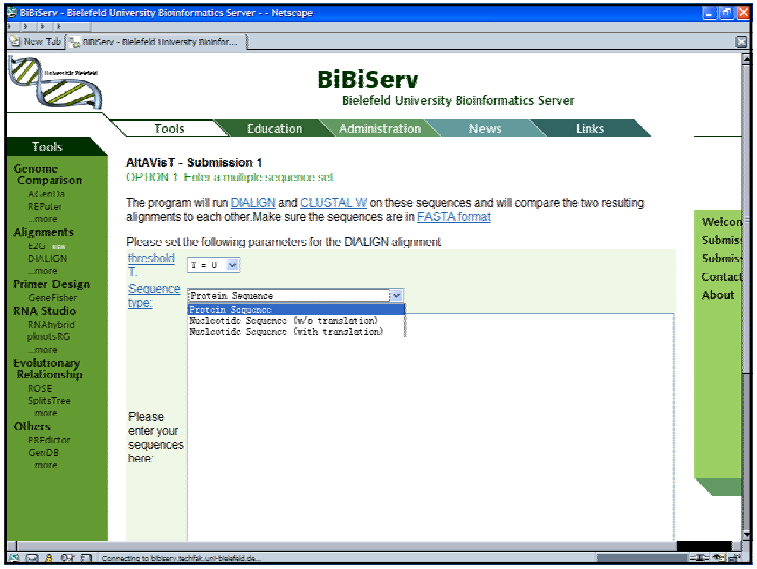
- **Altavist**
- **CHAOS + DIALIGN**

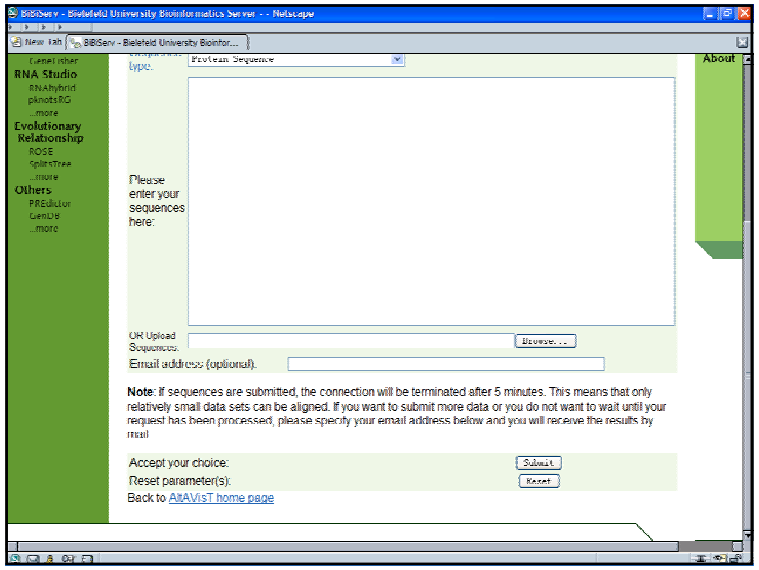
Altavist

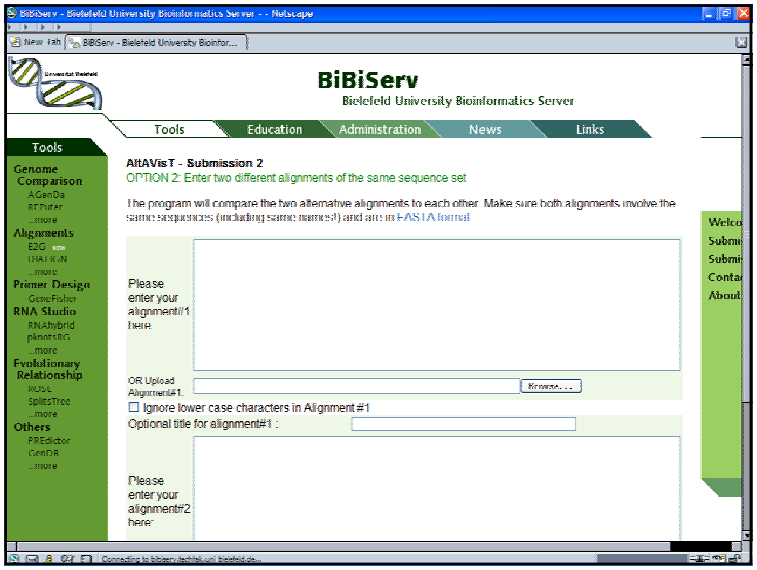
- AltAVist (Alternative Alignment Visualization Tool) is a WWW-based software program that is able to compare two alternative multiple alignments of a given sequence set to each other. Regions where both alignments coincide are color-coded to visualize the local agreement between the two alignments and to identify those regions of the alignments that can be considered to be most reliable.

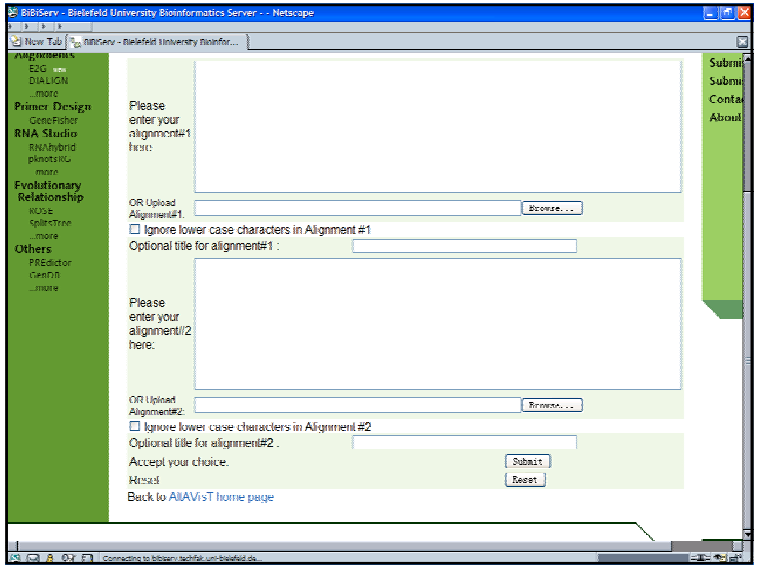


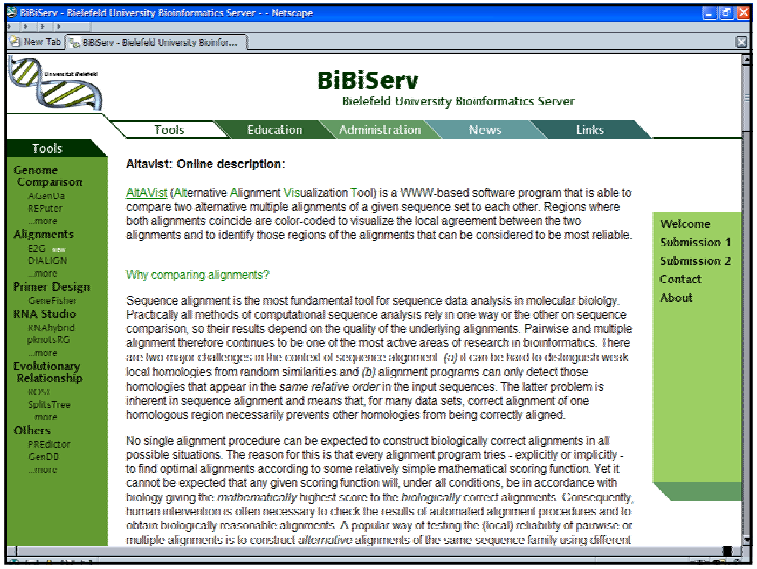


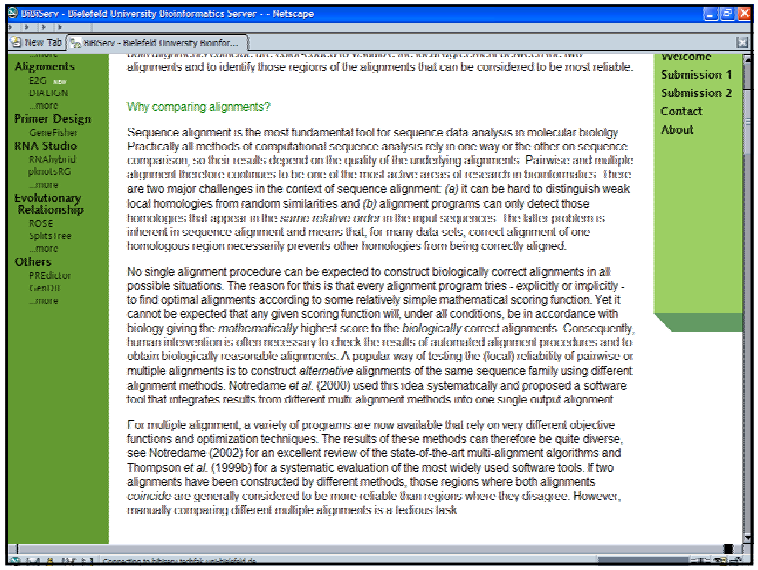


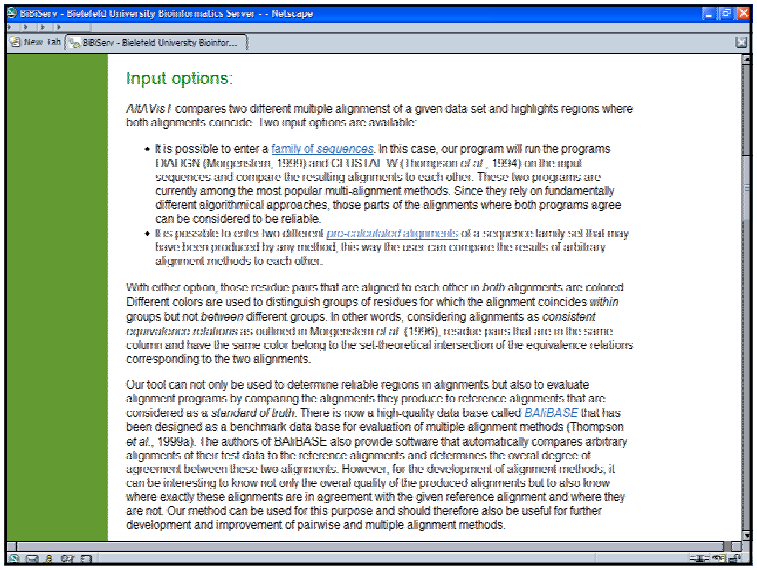

BiBiServ - Bielefeld University Bioinformatics Server - - Netscape
RNA Studio
RNAhybrid
pknotsRG
Evolutionary Relationship
ROSE
SplitsTree
Others
PREdictor
GenDB
Please enter your sequences here:
OR Upload Sequences:
Browse...
Email address (optional):
Note: If sequences are submitted, the connection will be terminated after 5 minutes. This means that only relatively small data sets can be aligned. If you want to submit more data or you do not want to wait until your request has been processed, please specify your email address below and you will receive the results by mail
Accept your choice:
Submit
Reset parameter(s):
Reset
Back to AltAVisT home page
About

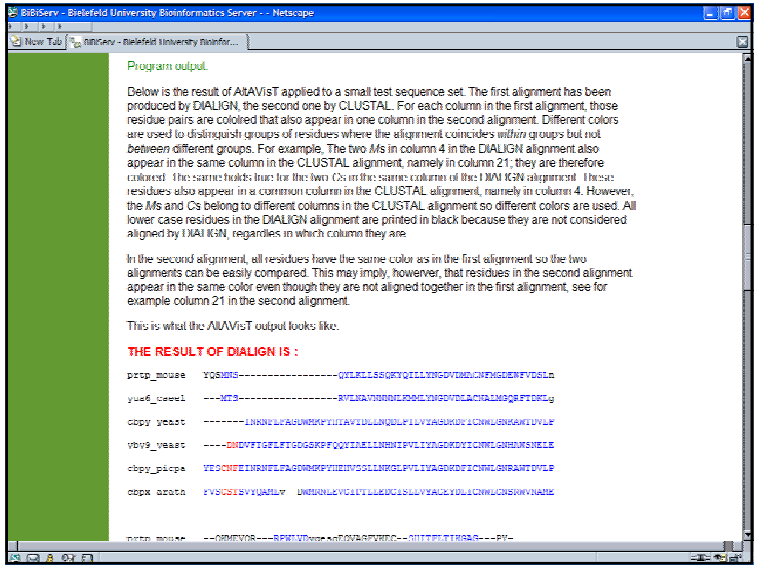

BiBiServ
Bielefeld University Bioinformatics Server
Tools
Education
Administration
News
Links
Genome Comparison
Alignments
Primer Design
RNA Studio
Evolutionary Relationship
Others
AltAVisT - Submission 2
OPTION 2: Enter two different alignments of the same sequence set
The program will compare the two alternative alignments to each other. Make sure both alignments involve the same sequences (including same names) and are in FASTA format
Please enter your alignment#1 here:
OR Upload Alignment#1:
Browse...
Ignore lower case characters in Alignment #1
Optional title for alignment#1 :
Please enter your alignment#2 here:

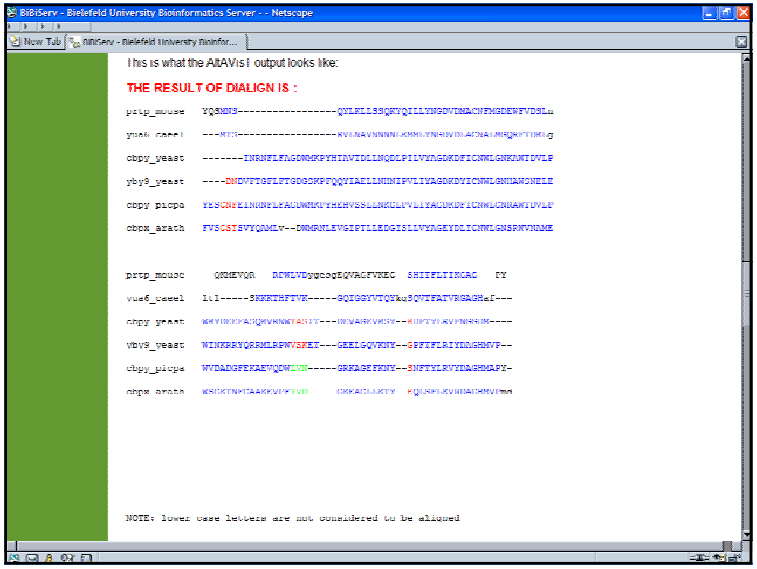

BiBiServ - Bielefeld University Bioinformatics Server - - Netscape
Primer Design
GeneFisher
RNA Studio
RNAhybrid
pknotsRG
Evolutionary Relationship
ROSE
SplitsTree
Others
PREdictor
GenDB
Please enter your alignment#1 here:
OR Upload Alignment#1:
Browse...
Ignore lower case characters in Alignment #1
Optional title for alignment#1 :
Please enter your alignment#2 here:
OR Upload Alignment#2:
Browse...
Ignore lower case characters in Alignment #2
Optional title for alignment#2 :
Accept your choice:
Submit
Reset
Back to AltAVisT home page
About

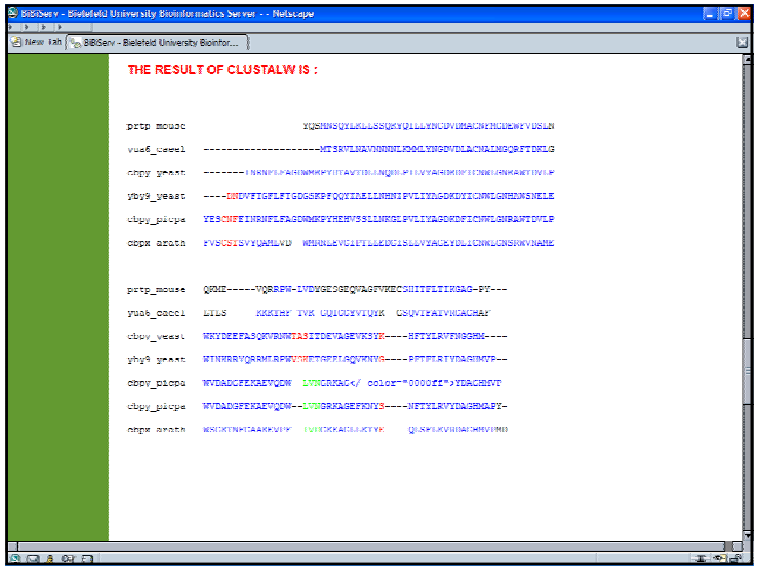

BiBiServ
Bielefeld University Bioinformatics Server
Tools
Education
Administration
News
Links
Genome Comparison
Alignments
Primer Design
RNA Studio
Evolutionary Relationship
Others
Altavist: Online description:
AltAVist (Alternative Alignment Visualization Tool) is a WWW-based software program that is able to compare two alternative multiple alignments of a given sequence set to each other. Regions where both alignments coincide are color-coded to visualize the local agreement between the two alignments and to identify those regions of the alignments that can be considered to be most reliable.
Why comparing alignments?
Sequence alignment is the most fundamental tool for sequence data analysis in molecular biology. Practically all methods of computational sequence analysis rely in one way or the other on sequence comparison, so their results depend on the quality of the underlying alignments. Pairwise and multiple alignment therefore continues to be one of the most active areas of research in bioinformatics. There are two major challenges in the control of sequence alignment: (a) it can be hard to distinguish weak local homologies from random similarities and (b) alignment programs can only detect those homologies that appear in the same relative order in the input sequences. The latter problem is inherent in sequence alignment and means that, for many data sets, correct alignment of one homologous region necessarily prevents other homologies from being correctly aligned.
No single alignment procedure can be expected to construct biologically correct alignments in all possible situations. The reason for this is that every alignment program tries - explicitly or implicitly - to find optimal alignments according to some relatively simple mathematical scoring function. Yet it cannot be expected that any given scoring function will, under all conditions, be in accordance with biology giving the mathematically highest score to the biologically correct alignments. Consequently, human intervention is often necessary to check the results of automated alignment procedures and to obtain biologically reasonable alignments. A popular way of testing the (local) reliability of pairwise or multiple alignments is to construct alternative alignments of the same sequence family using different
Welcome
Submission 1
Submission 2
Contact
About

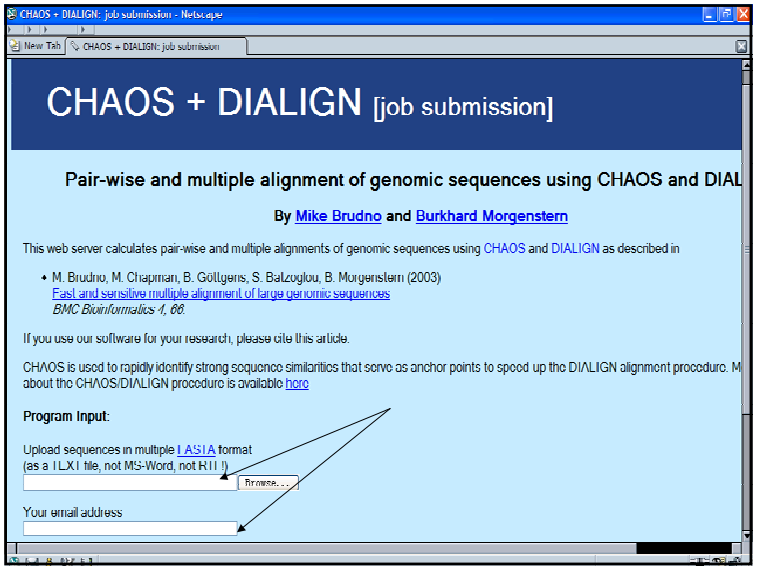

BiBiServ - Bielefeld University Bioinformatics Server - - Netscape
Alignments
Primer Design
RNA Studio
Evolutionary Relationship
Others
Welcome
Submission 1
Submission 2
Contact
About
alignments and to identify those regions of the alignments that can be considered to be most reliable.
Why comparing alignments?
Sequence alignment is the most fundamental tool for sequence data analysis in molecular biology. Practically all methods of computational sequence analysis rely in one way or the other on sequence comparison, so their results depend on the quality of the underlying alignments. Pairwise and multiple alignment therefore continues to be one of the most active areas of research in bioinformatics. There are two major challenges in the context of sequence alignment: (a) it can be hard to distinguish weak local homologies from random similarities and (b) alignment programs can only detect those homologies that appear in the same relative order in the input sequences. The latter problem is inherent in sequence alignment and means that, for many data sets, correct alignment of one homologous region necessarily prevents other homologies from being correctly aligned.
No single alignment procedure can be expected to construct biologically correct alignments in all possible situations. The reason for this is that every alignment program tries - explicitly or implicitly - to find optimal alignments according to some relatively simple mathematical scoring function. Yet it cannot be expected that any given scoring function will, under all conditions, be in accordance with biology giving the mathematically highest score to the biologically correct alignments. Consequently, human interventions is often necessary to check the results of automated alignment procedures and to obtain biologically reasonable alignments. A popular way of testing the (local) reliability of pairwise or multiple alignments is to construct alternative alignments of the same sequence family using different alignment methods. Notredame et al. (2000) used this idea systematically and proposed a software tool that integrates results from different multi alignment methods into one single output alignment.
For multiple alignment, a variety of programs are now available that rely on very different objective functions and optimization techniques. The results of these methods can therefore be quite diverse, see Notredame (2002) for an excellent review of the state-of-the-art multi-alignment algorithms and Thompson et al. (1999b) for a systematic evaluation of the most widely used software tools. If two alignments have been constructed by different methods, those regions where both alignments coincide are generally considered to be more reliable than regions where they disagree. However, manually comparing different multiple alignments is a tedious task.

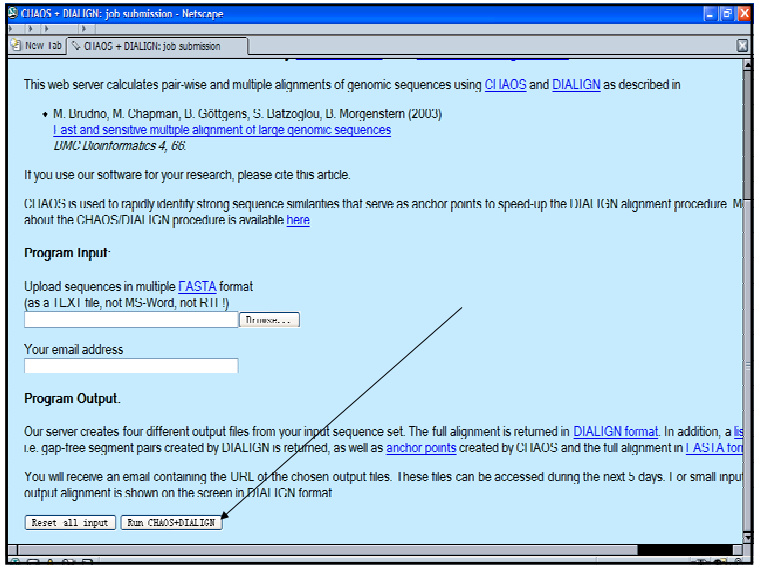

BiBiServ - Bielefeld University Bioinformatics Server - - Netscape
Input options:
AltAVisT compares two different multiple alignmenst of a given data set and highlights regions where both alignments coincide. Two input options are available:
It is possible to enter a family of sequences. In this case, our program will run the programs DIALIGN (Morgenstern, 1999) and CLUSTAL W (Thompson et al., 1994) on the input sequences and compare the resulting alignments to each other. These two programs are currently among the most popular multi-alignment methods. Since they rely on fundamentally different algorithmical approaches, those parts of the alignments where both programs agree can be considered to be reliable.
It is possible to enter two different pre-calculated alignments of a sequence family set that may have been produced by any method, this way the user can compare the results of arbitrary alignment methods to each other.
With either option, those residue pairs that are aligned to each other in both alignments are colored. Different colors are used to distinguish groups of residues for which the alignment coincides within groups but not between different groups. In other words, considering alignments as consistent equivalence relations as outlined in Morgenstern et al. (1996), residue pairs that are in the same column and have the same color belong to the set-theoretical intersection of the equivalence relations corresponding to the two alignments.
Our tool can not only be used to determine reliable regions in alignments but also to evaluate alignment programs by comparing the alignments they produce to reference alignments that are considered as a standard of truth. There is now a high-quality data base called BAliBASE that has been designed as a benchmark data base for evaluation of multiple alignment methods (Thompson et al., 1999a). The authors of BAliBASE also provide software that automatically compares arbitrary alignments of their test data to the reference alignments and determines the overall degree of agreement between these two alignments. However, for the development of alignment methods, it can be interesting to know not only the overal quality of the produced alignments but to also know where exactly these alignments are in agreement with the given reference alignment and where they are not. Our method can be used for this purpose and should therefore also be useful for further development and improvement of pairwise and multiple alignment methods.

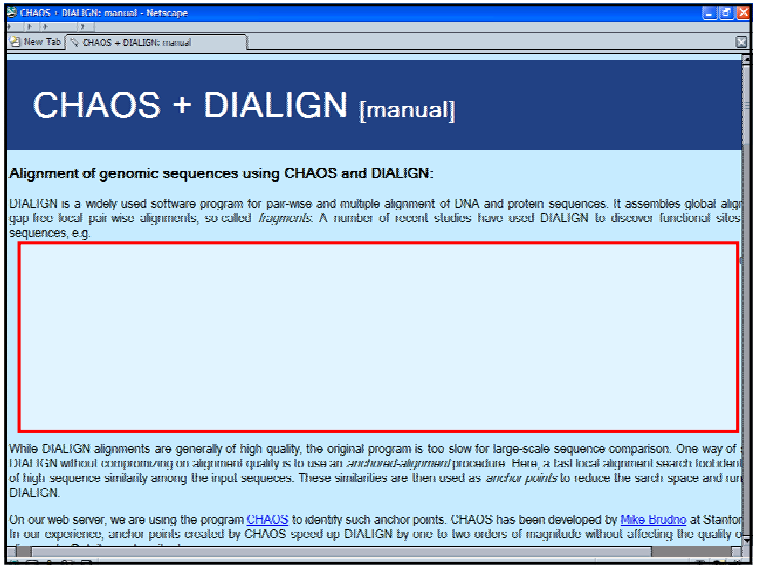

BiBiServ - Bielefeld University Bioinformatics Server - - Netscape
Program output
Below is the result of AltAVisT applied to a small test sequence set. The first alignment has been produced by DIALIGN, the second one by CLUSTAL. For each column in the first alignment, those residue pairs are colored that also appear in one column in the second alignment. Different colors are used to distinguish groups of residues where the alignment coincides within groups but not between different groups. For example, The two Ms in column 4 in the DIALIGN alignment also appear in the same column in the CLUSTAL alignment, namely in column 21; they are therefore colored. The same holds true for the two Cs in the same column of the DIALIGN alignment. These residues also appear in a common column in the CLUSTAL alignment, namely in column 4. However, the Ms and Cs belong to different columns in the CLUSTAL alignment so different colors are used. All lower case residues in the DIALIGN alignment are printed in black because they are not considered aligned by DIALIGN, regardles in which column they are.
In the second alignment, all residues have the same color as in the first alignment so the two alignments can be easily compared. This may imply, however, that residues in the second alignment appear in the same color even though they are not aligned together in the first alignment, see for example column 21 in the second alignment.
This is what the AltAVisT output looks like:
THE RESULT OF DIALIGN IS :

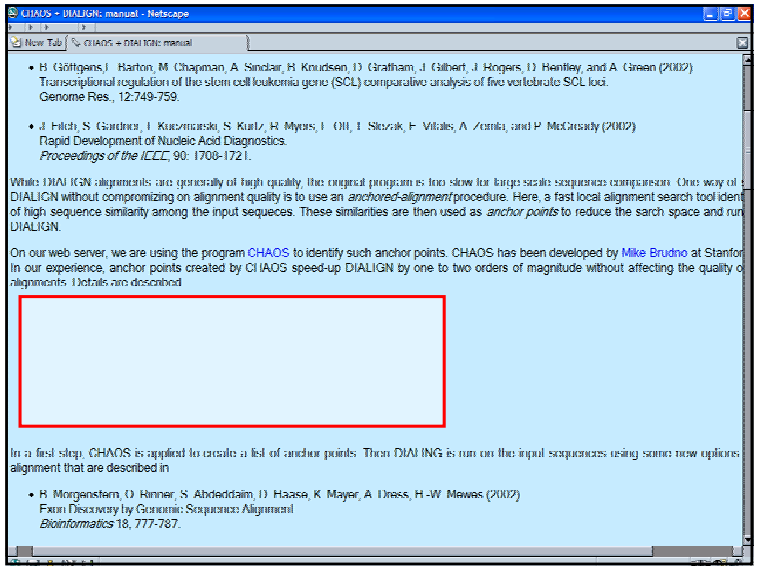

BiBiServ - Bielefeld University Bioinformatics Server - - Netscape
This is what the AltAVisT output looks like:
THE RESULT OF DIALIGN IS :
NOTE: lower case letters are not considered to be aligned

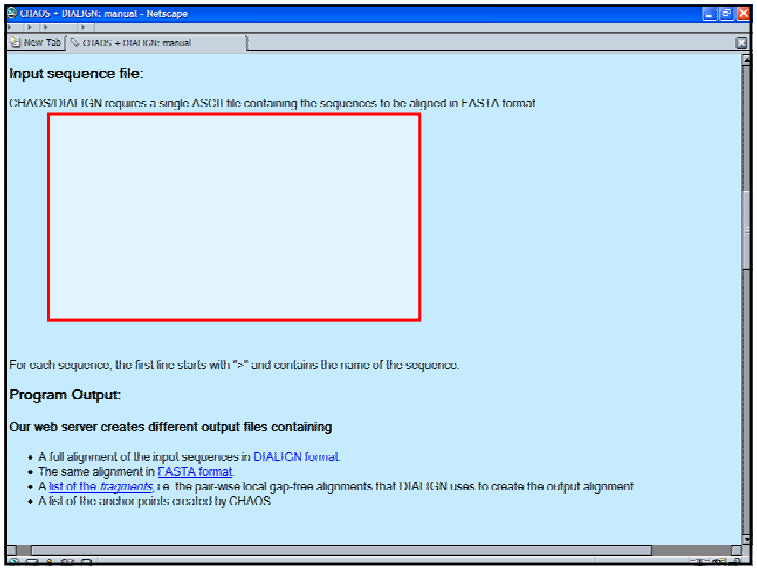

THE RESULT OF CLUSTALW IS :

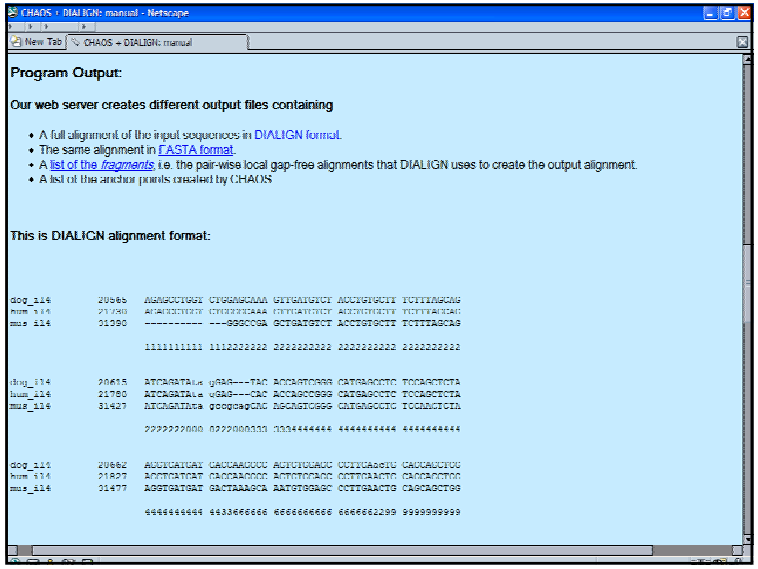

CHAOS + DIALIGN [job submission]
Pair-wise and multiple alignment of genomic sequences using CHAOS and DIAL
By Mike Brudno and Burkhard Morgenstern
This web server calculates pair-wise and multiple alignments of genomic sequences using CHAOS and DIALIGN as described in
M. Brudno, M. Chapman, B. Göttgens, S. Batzoglou, B. Morgenstern (2003)
Fast and sensitive multiple alignment of large genomic sequences
BMC Bioinformatics 4, 66.
If you use our software for your research, please cite this article.
CHAOS is used to rapidly identify strong sequence similarities that serve as anchor points to speed up the DIALIGN alignment procedure. M
about the CHAOS/DIALIGN procedure is available here
Program Input:
Upload sequences in multiple FASTA format
(as a TEXT file, not MS-Word, not RTF!)
Browse...
Your email address

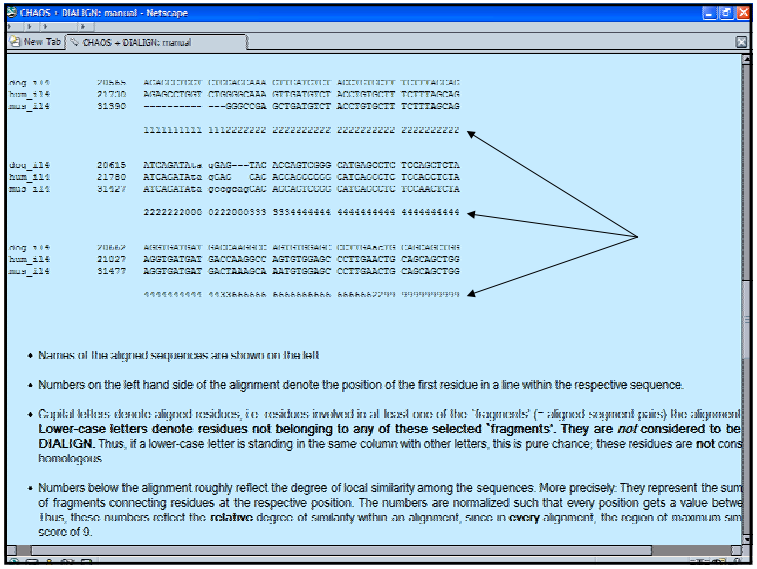

This web server calculates pair-wise and multiple alignments of genomic sequences using CHAOS and DIALIGN as described in
M. Brudno, M. Chapman, B. Göttgens, S. Batzoglou, B. Morgenstern (2003)
Fast and sensitive multiple alignment of large genomic sequences
BMC Bioinformatics 4, 66.
If you use our software for your research, please cite this article.
CHAOS is used to rapidly identify strong sequence similarities that serve as anchor points to speed-up the DIALIGN alignment procedure. M
about the CHAOS/DIALIGN procedure is available here
Program Input:
Upload sequences in multiple FASTA format
(as a TEXT file, not MS-Word, not RTF!)
Browse...
Your email address
Program Output:
Our server creates four different output files from your input sequence set. The full alignment is returned in DIALIGN format. In addition, a lis
i.e. gap-free segment pairs created by DIALIGN is returned, as well as anchor points created by CHAOS and the full alignment in FASTA for
You will receive an email containing the URL of the chosen output files. These files can be accessed during the next 5 days. For small input
output alignment is shown on the screen in DIALIGN format
Reset all input
Run CHAOS+DIALIGN

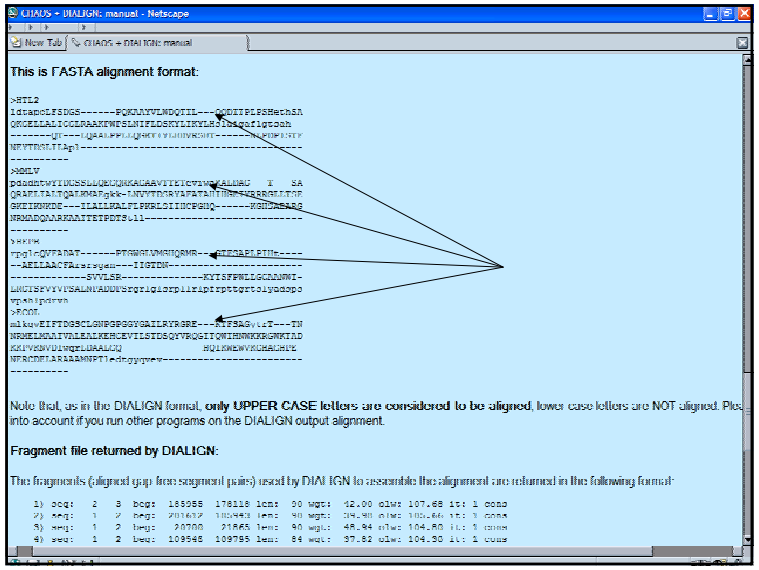

CHAOS + DIALIGN [manual]
Alignment of genomic sequences using CHAOS and DIALIGN:
DIALIGN is a widely used software program for pair-wise and multiple alignment of DNA and protein sequences. It assembles global align
gap free local pair wise alignments, so called fragments. A number of recent studies have used DIALIGN to discover functional sites
sequences, e.g.
While DIALIGN alignments are generally of high quality, the original program is too slow for large-scale sequence comparison. One way of
DIALIGN without compromising on alignment quality is to use an anchored-alignment procedure. Here, a fast local alignment search tool identi
of high sequence similarity among the input sequences. These similarities are then used as anchor points to reduce the search space and run
DIALIGN.
On our web server, we are using the program CHAOS to identify such anchor points. CHAOS has been developed by Mike Brudno at Stanfor
In our experience, anchor points created by CHAOS speed up DIALIGN by one to two orders of magnitude without affecting the quality o

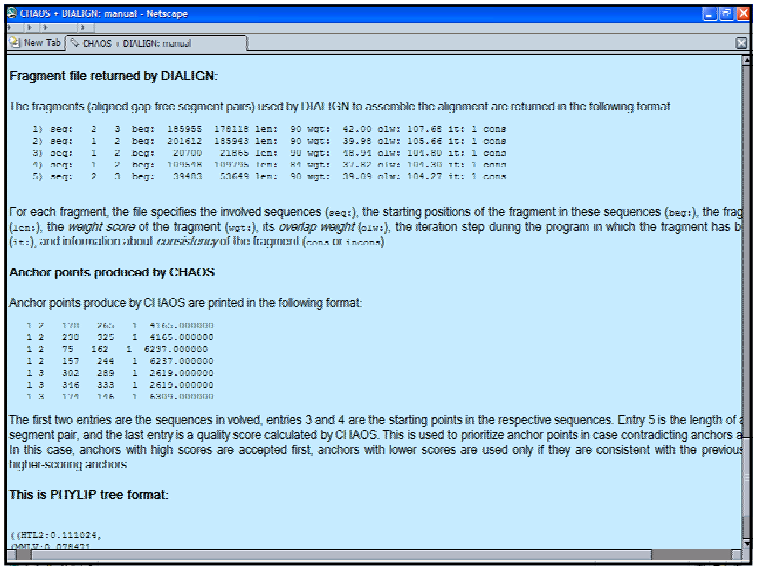

CHAOS + DIALIGN: manual - Netscape
New Tab
CHAOS + DIALIGN: manual
B Göttgens, L Barton, M Chapman, A Sinclair, R Knudsen, D Grafham, J Gilbert, J Rogers, D Bentley, and A Green (2002)
Transcriptional regulation of the stem cell leukemia gene (SCL) comparative analysis of five vertebrate SCL loci.
Genome Res., 12:749-759.
J Fitch, S Gardner, L Kaczmarski, S Kurtz, R Myers, L Ott, I Slezak, F Vilalis, A Zemla, and P McGready (2002)
Rapid Development of Nucleic Acid Diagnostics.
Proceedings of the ICCC, 90: 1708-1721.
DIALIGN
DIALIGN
In a first step, CHAOS is applied to create a list of anchor points. Then DIALIGN is run on the input sequences using some new options
alignment that are described in
B Morgenstern, O Rinner, S Abdeddaim, D Haase, K Mayer, A Dress, H-W Mewes (2002)
Exon Discovery by Genomic Sequence Alignment
Bioinformatics 18, 777-787.

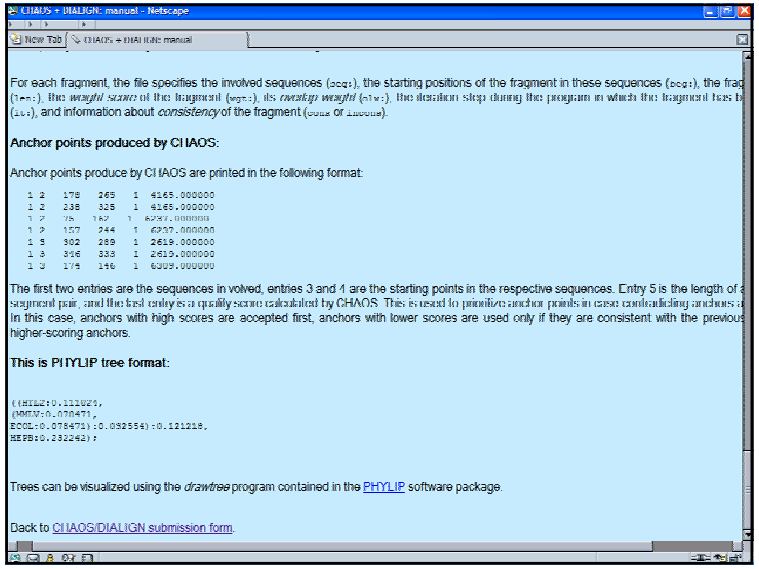

CHAOS + DIALIGN: manual - Netscape
New Tab
CHAOS + DIALIGN: manual
Input sequence file:
CHAOS/DIALIGN requires a single ASCII file containing the sequences to be aligned in FASTA format
For each sequence, the first line starts with ">" and contains the name of the sequence.
Program Output:
Our web server creates different output files containing
A full alignment of the input sequences in DIALIGN format.
The same alignment in FASTA format.
A list of the fragments, i.e. the pair-wise local gap-free alignments that DIALIGN uses to create the output alignment
A list of the anchor points created by CHAOS

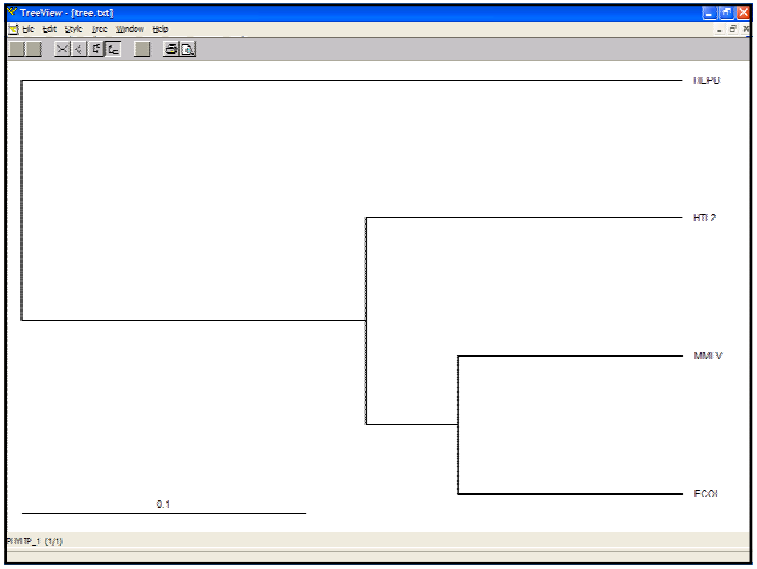

CHAOS + DIALIGN: manual - Netscape
New Tab
CHAOS + DIALIGN: manual
Program Output:
Our web server creates different output files containing
A full alignment of the input sequences in DIALIGN format.
The same alignment in FASTA format.
A list of the fragments, i.e. the pair-wise local gap-free alignments that DIALIGN uses to create the output alignment.
A list of the anchor points created by CHAOS
This is DIALIGN alignment format:


CHAOS + DIALIGN: manual - Netscape
New Tab
CHAOS + DIALIGN: manual
Names of the aligned sequences are shown on the left
Numbers on the left hand side of the alignment denote the position of the first residue in a line within the respective sequence.
Capital letters denote aligned residues, i.e. residues involved in at least one of the "fragments" (= aligned segment pairs) the alignment
Lower-case letters denote residues not belonging to any of these selected "fragments". They are not considered to be
DIALIGN. Thus, if a lower-case letter is standing in the same column with other letters, this is pure chance; these residues are not cons
homologous
Numbers below the alignment roughly reflect the degree of local similarity among the sequences. More precisely: They represent the sum
of fragments connecting residues at the respective position. The numbers are normalized such that every position gets a value betwe
Thus, these numbers reflect the relative degree of similarity within an alignment, since in every alignment, the region of maximum sim
score of 9.


CHAOS + DIALIGN: manual - Netscape
This is FASTA alignment format:
Note that, as in the DIALIGN format, only UPPER CASE letters are considered to be aligned, lower case letters are NOT aligned. Pla
into account if you run other programs on the DIALIGN output alignment.
Fragment file returned by DIALIGN:
The fragments (aligned gap free segment pairs) used by DIALIGN to assemble the alignment are returned in the following format:


CHAOS + DIALIGN: manual - Netscape
Fragment file returned by DIALIGN:
The fragments (aligned gap free segment pairs) used by DIALIGN to assemble the alignment are returned in the following format:
Anchor points produced by CHAOS
Anchor points produce by CHAOS are printed in the following format:
The first two entries are the sequences in volved, entries 3 and 4 are the starting points in the respective sequences. Entry 5 is the length of a
segment pair, and the last entry is a quality score calculated by CHAOS. This is used to prioritize anchor points in case contradicting anchors a
In this case, anchors with high scores are accepted first, anchors with lower scores are used only if they are consistent with the previous
higher-scoring anchors
This is PHYLIP tree format:


CHAOS + DIALIGN: manual - Netscape
Anchor points produced by CHAOS:
Anchor points produce by CHAOS are printed in the following format:
1 2 178 265 1 4165.000000
1 2 238 325 1 4165.000000
1 2 157 244 1 6237.000000
1 3 302 289 1 2619.000000
1 3 316 333 1 2619.000000
1 3 174 146 1 6309.000000
The first two entries are the sequences in volved, entries 3 and 4 are the starting points in the respective sequences. Entry 5 is the length of a
segment pair, and the last entry is a quality score calculated by CHAOS. This is used to prioritize anchor points in case contradicting anchors a
In this case, anchors with high scores are accepted first, anchors with lower scores are used only if they are consistent with the previous
higher-scoring anchors.
This is PHYLIP tree format:
((HTL2:0.111024,
(MMLV:0.070471,
ECOL:0.078471):0.092554):0.121218,
HEPB:0.232242);
Trees can be visualized using the drawtree program contained in the PHYLIP software package.
Back to CHAOS/DIALIGN submission form.


TreeView - [tree.txt]
HEPB
HTL2
MMLV
ECOL
0.1