

alinhamento global-
alinhamento múltiplo de
seqüências

Alinhamento múltiplos de seqüências

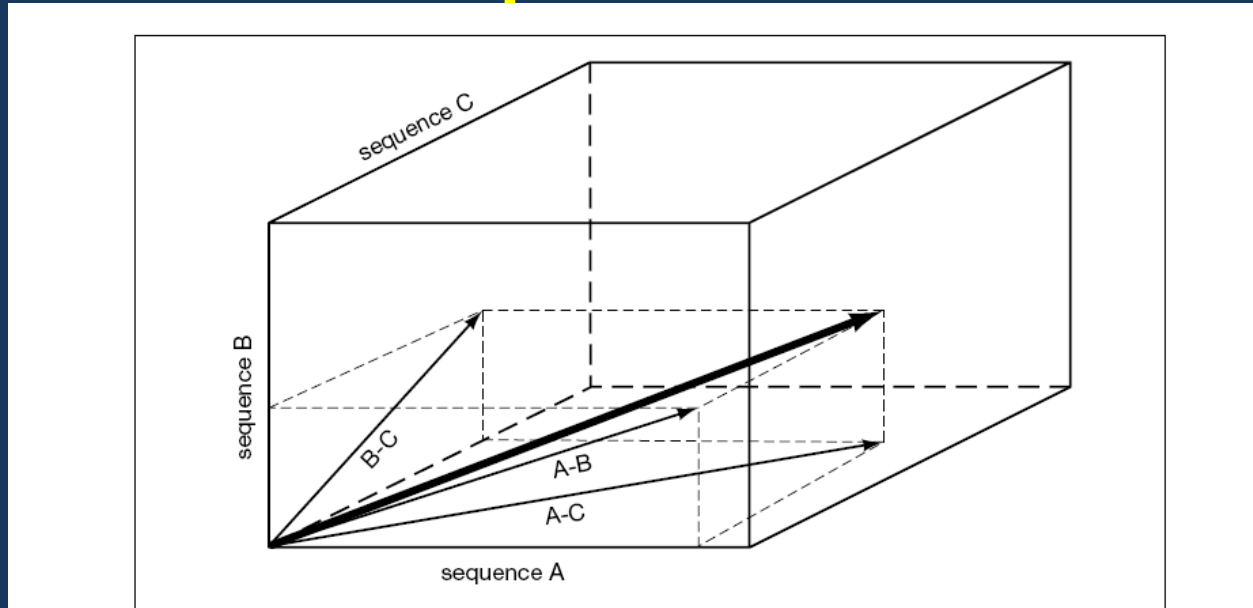
Qual a importância de se realizar alinhamentos múltiplos em oposição a alinhamentos em pares?

Alinhamento múltiplos de seqüências

Usos do alinhamento múltiplo de seqüências

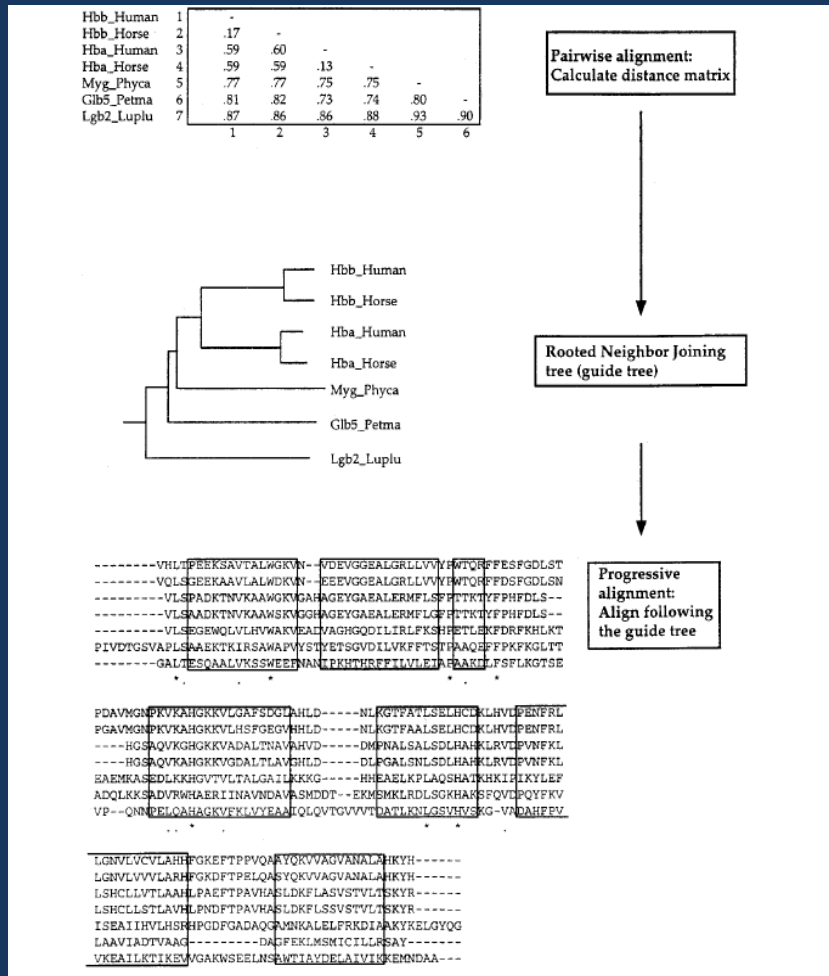
- Estabelecer relações filogenéticas entre diferentes proteínas
- Predizer domínios conservados, incluindo resíduos críticos para a função da proteína
- Comparar a proteína de interesse de forma mais detalhada com membros da mesma família de proteínas

Alinhamento múltiplo de seqüências



Ao se alinhar 3 seqüências pode-se calcular o alinhamento ótimo através do dynamic programming, entretanto neste caso você passa a ter que calcular uma matriz tridimensional. Caso mais seqüências fossem adicionada seria necessário uma matriz N dimensional onde N é o numero de seqüências alinhadas. Isso cria um problema computacional pois o numero de comparações a ser feitas será o numero de aminoácidos elevado ao numero de seqüências alinhadas.

Alinhamento progressivo (Clustal)



- Com a finalidade de solucionar este problema um método heurístico foi desenvolvido de modo a fazer um alinhamento progressivo de seqüências baseado na sua proximidade em uma árvore filogenética baseada em alinhamentos par a par
- Apesar deste tipo de método gerar bons alinhamentos em tempo razoável, ele não garante que o melhor alinhamento possível será obtido

Alinhamento progressivo (Clustal)

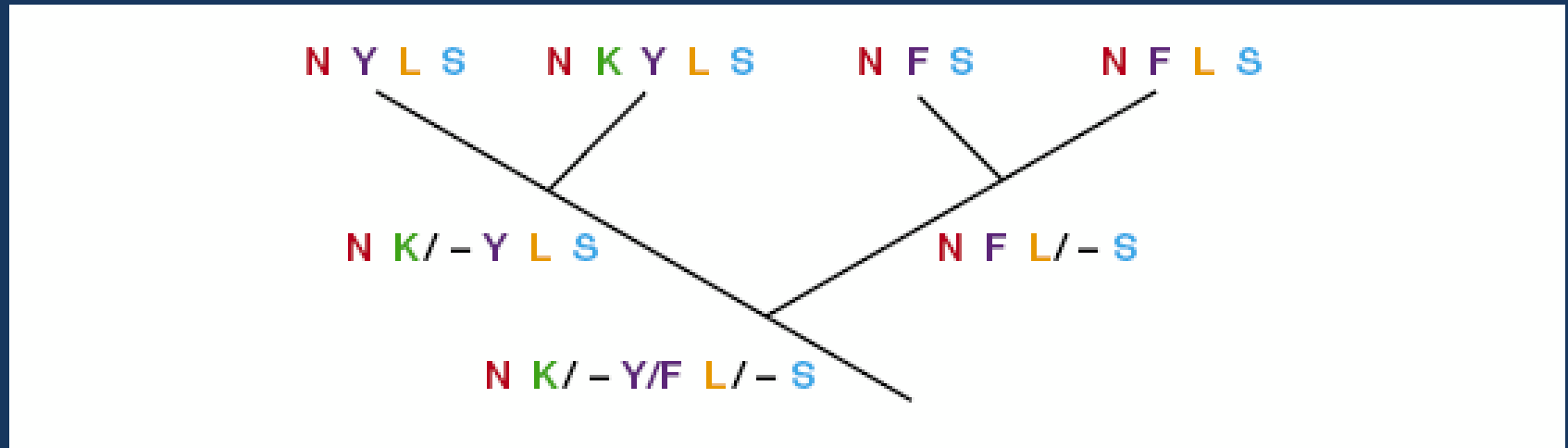
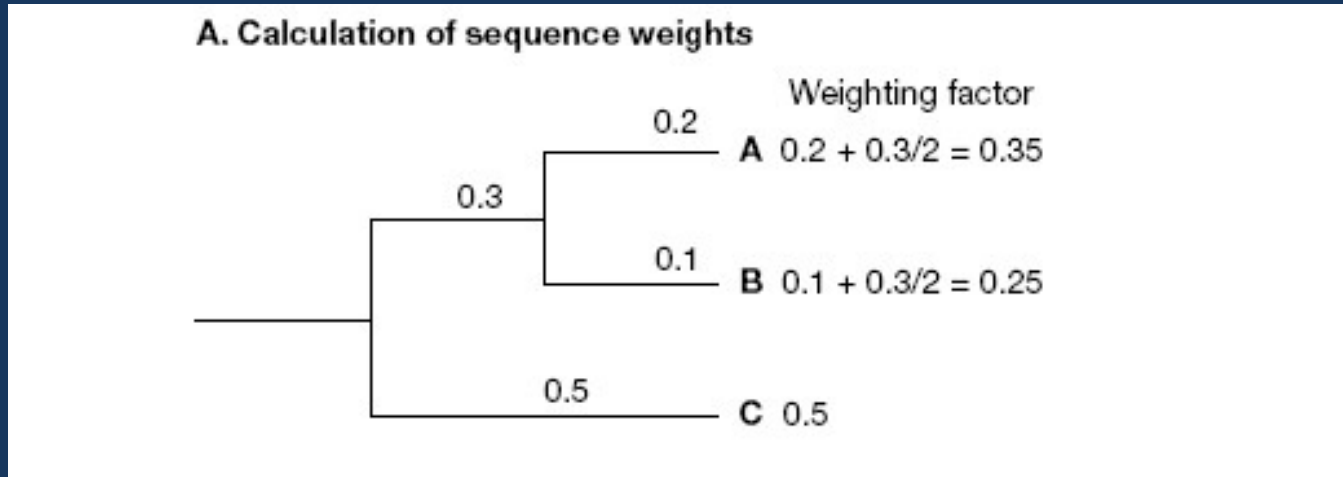


Figura mostra o princípio do alinhamento progressivo. Alinhamentos são realizados seqüencialmente baseado na árvore filogenética estimada criada a partir dos alinhamentos par a par realizados.

Alinhamento progressivo (Clustal)



Ao realizar cada alinhamento são gerados pesos para cada seqüências dependendo da sua distancia em relação as outras seqüências do alinhamento. O objetivo deste procedimento é evitar a criação de vieses gerados pela escolhas de muitas seqüências próximas em oposição a poucas seqüências divergentes.

Alinhamento progressivo (Clustal)

B. Use of sequence weights

Column in alignment 1

Sequence A (weight a)K.....

Sequence B (weight b)I.....

Column in alignment 2

Sequence C (weight c)L.....

Sequence D (weight d)V.....

Score for matching these two column in an msa =

$$\begin{aligned} & [a \times c \times \text{score}(K,L) + \\ & a \times d \times \text{score}(K,V) + \\ & b \times c \times \text{score}(I,L) + \\ & b \times d \times \text{score}(I,V)] / 4 \end{aligned}$$

Exemplo de calculo de escore em alinhamento progressivo a partir de dois pares de seqüências alinhadas. Note que cada seqüência possui um peso especifico mesmo já estando alinhada com uma outra seqüência. Escores de substituição ou identidade entre seqüências já alinhadas (Seq A e B, no exemplo acima) não são computados visto que o alinhamento entre A e B não vai mais variar.

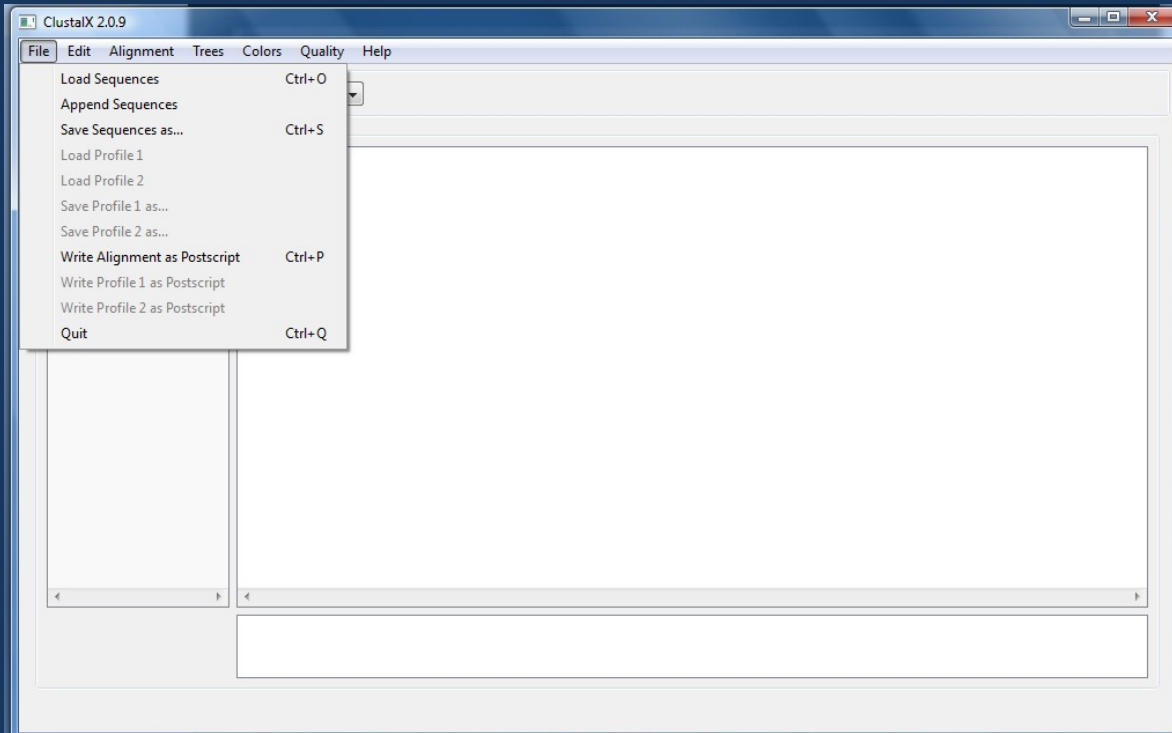
Problemas do método de alinhamento progressivo

- Este tipo de algoritmo é altamente dependente da qualidade dos alinhamentos iniciais, visto que ao longo do processo eles não serão mais alterados
- Dificuldade é maior quando mesmo as seqüências mais próximas são distantemente relacionadas
- Dificuldade em estabelecer parâmetro ótimos de alinhamento

Uso de aproximação sucessiva para refinar alinhamentos

- Devido a limitações impostas devido ao alinhamento progressivo foi implementado nas ultimas versões do clustal um algoritmo utilizando aproximação sucessiva (iteration) para minimizar este tipo de problema
- O algoritmo realiza após cada alinhamento um procedimento no qual ele seleciona uma seqüência e realinha esta com o resto do alinhamento, caso o escore resultante for melhor que o inicial o novo alinhamento é mantido. Isto é realizado sucessivamente com todas as seqüências do alinhamento.

ClustalX



- Seqüências em formato multi-fasta ou aln pode ser carregadas no clustalX através do comando “load sequences”

ClustalX

Cada linha representa uma sequência

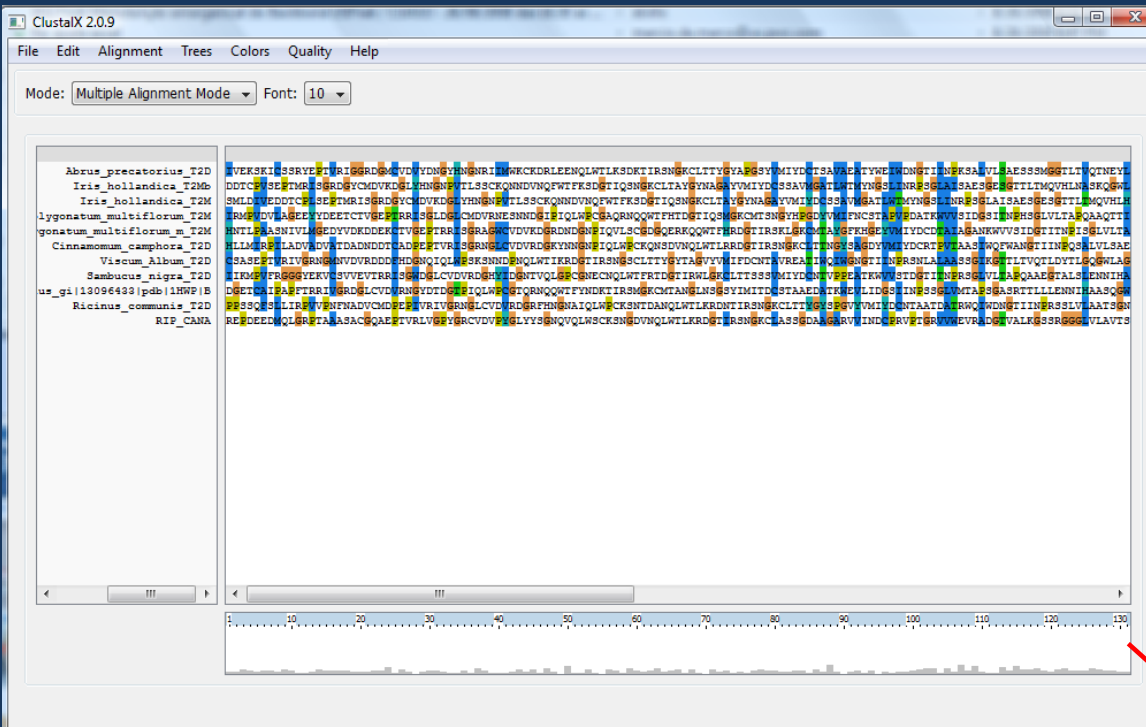


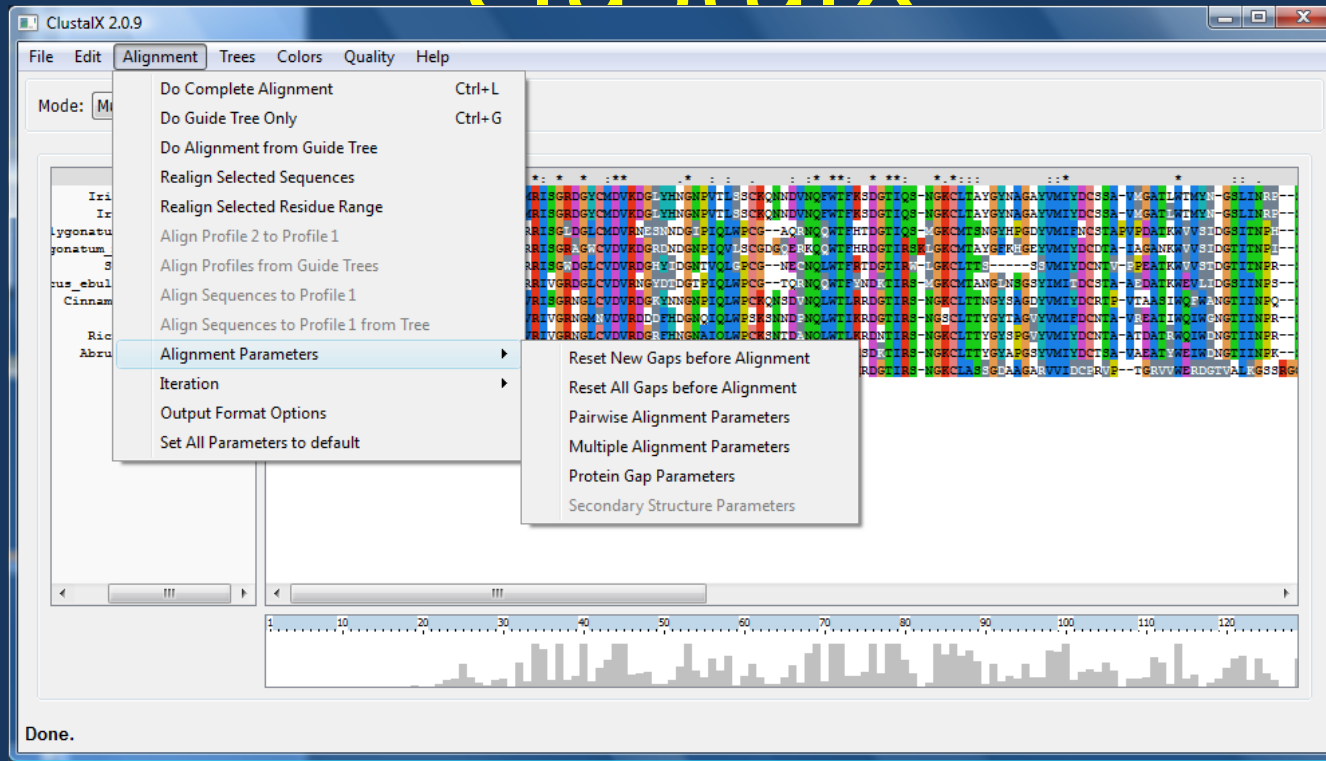
Gráfico com Indicação do grau de conservação

Sistema de cores do clustal

ClustalX	Rules:
blue	(W,L,V,I,M,F): {50%, p}{60%, wlvimafcyhp} (A): {50%, p}{60%, wlvimafcyhp}{85%, t,s,g} (C): {50%, p}{60%, wlvimafcyhp}{85%, s}
red	(K,R): {60%, kr}{85%, q}
green	(T): {50%, ts}{60%, wlvimafcyhp} (S): {50%, ts}{80%, wlvimafcyhp} (N): {50%, n}{85%, d} (Q): {50%, qe}{60%, kr}
pink	(C): {85%, c}
magenta	(D): {50%, de,n} (E): {50%, de,qe}
orange	(G): {always}
cyan	(H,Y): {50%, p}{60%, wlvimafcyhp}
yellow	(P): {always}

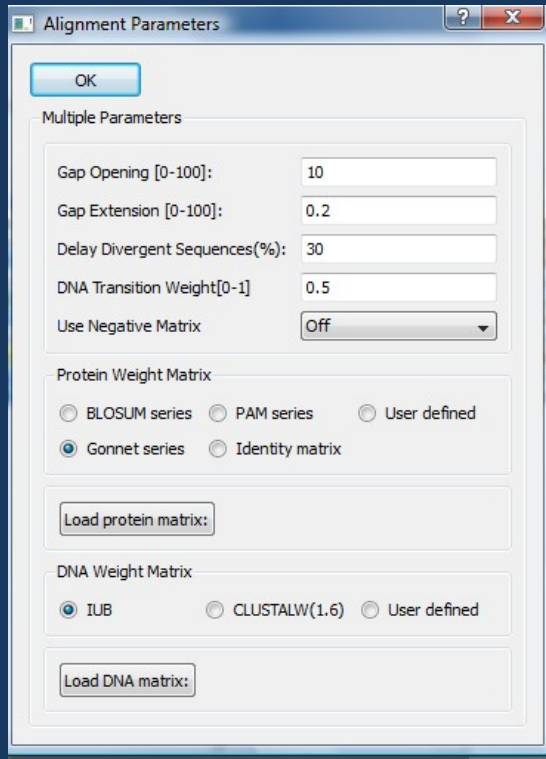
- O fundo colorido do clustal tem duas funções: chamar a atenção de regiões conservadas e para o caráter dos aminoácidos

Alinhando seqüências no clustalX



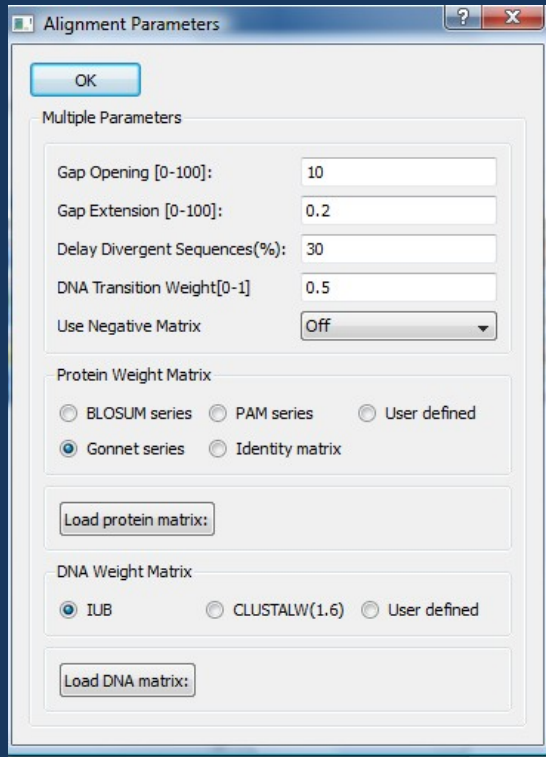
Menu “alignment” permite acessar diversos comandos incluindo menus de ajuste de parâmetros de alinhamentos

Alinhando seqüências no clustalX



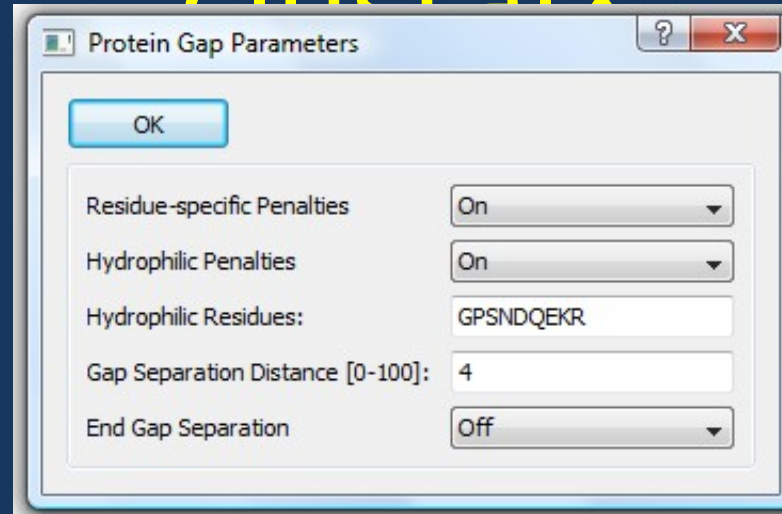
- Clicando no item “multiple alignment parameters” é possível acessar um menu onde é possível ajustar os seguintes parametros
- Gap opening- Penalização de escore para iniciar uma região de “gap”
- Gap extension- Penalização de escore para estender uma região de “gap” (normalmente menor que o Gap opening)
- Delay divergent sequences- Atraso de alinhamento de seqüências divergentes que somente serão alinhadas após as outras sequencias (porcentagem de identidade abaixo da qual que faz a seqüência seja considerada divergente)

Alinhando seqüências no clustalX



- Transition weight (somente DNA)- da a transições (A<->G, T<->C) um escore diferente de 0.
- Use negative matrix- permite o uso de matrizes negativas, importante quando as seqüências forem relacionadas somente em uma pequena porção. Em condições normais prejudica um pouco o alinhamento
- Protein Weight Matrix- Matriz de substituição a ser utilizada (note que voce só escolhe a serie de matriz: Blosum, PAM, etc... O tipo de matriz dentro desta serie (por exemplo Blosum 62, blosum 80) é escolhido automaticamente pelo programa

Alinhando seqüências no clustalX



- Residue specific penalty- Considera a vizinhança de alguns resíduos como mais ou menos favoráveis para abertura de “gaps”
- Hydrophilic penalties- aumentqa a chance de gaps em regiões ricas em resíduos hidrofílicos, que usualmente representam regiões menos estruturadas.
- Hydrophilic residues- especifica que residuos são considerados hidrofílicos
- Gap separation distance- numero de resíduos de distancia de uma região com gap na qual é penalizada uma nova abertura de gap
- Considera o gap no final da seqüência normalmente para o parâmetro acima

Alinhando seqüências no clustalX

- Como realiza um alinhamento global o clustal tentará realizar o alinhamento de toda a proteína não “jogando fora pedaços” como ocorre no caso de alinhamento por blast. Em adição não há a geração de um parâmetro de confiança que permita você avaliar a significância do seu alinhamento
- Deste modo o clustal é uma ferramenta muito pobre para identificação de função de proteínas. O seu principal uso é para uma caracterização mais fina após a determinação da identidade e domínios presentes na seqüência.

Alinhando seqüências no clustalX

- É muito importante que haja uma seleção criteriosa de que seqüências e quais regiões destas seqüência serão alinhadas
- Muitas vezes ao invés de selecionarmos proteínas inteiras, que podem conter um mosaico de regiões com diversas origens evolutivas, é preferível alinhar somente regiões de domínios em comum

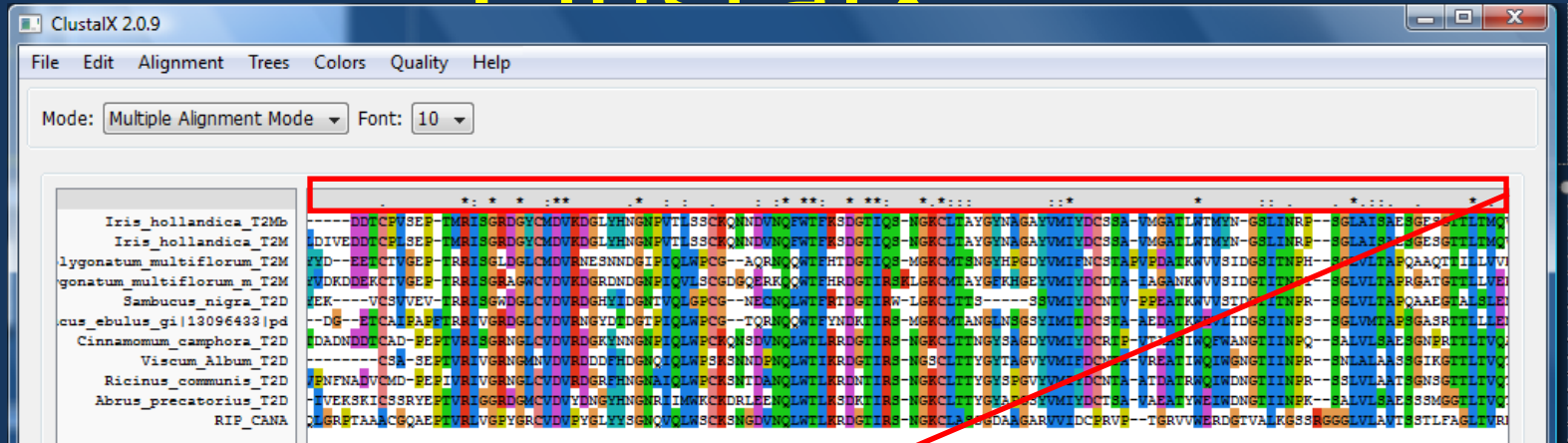
Alinhando seqüências no clustalX

- Alinhamento de múltiplas seqüências são bastante influenciados pelas penalizações de abertura e extensão de gaps e ao fazer alinhamentos o usuário normalmente deve ajustar estes parâmetros de modo a obter um bom alinhamento
- Para o alinhamento de seqüências mais divergentes é necessário a utilização de penalizações menores para a abertura de gaps

Alinhando seqüências no clustalX

- Devido aos problemas do algoritmo de alinhamento progressivo quando alinhamos seqüências distantes é sempre recomendável inserir no alinhamento múltiplas seqüências adicionais que sejam mais próximas das seqüências analisadas
- De modo geral , ter um numero razoável de seqüência ajuda o programa de alinhamento múltiplo e facilita a interpretação dos dados

Resultado de alinhamento ClustalX

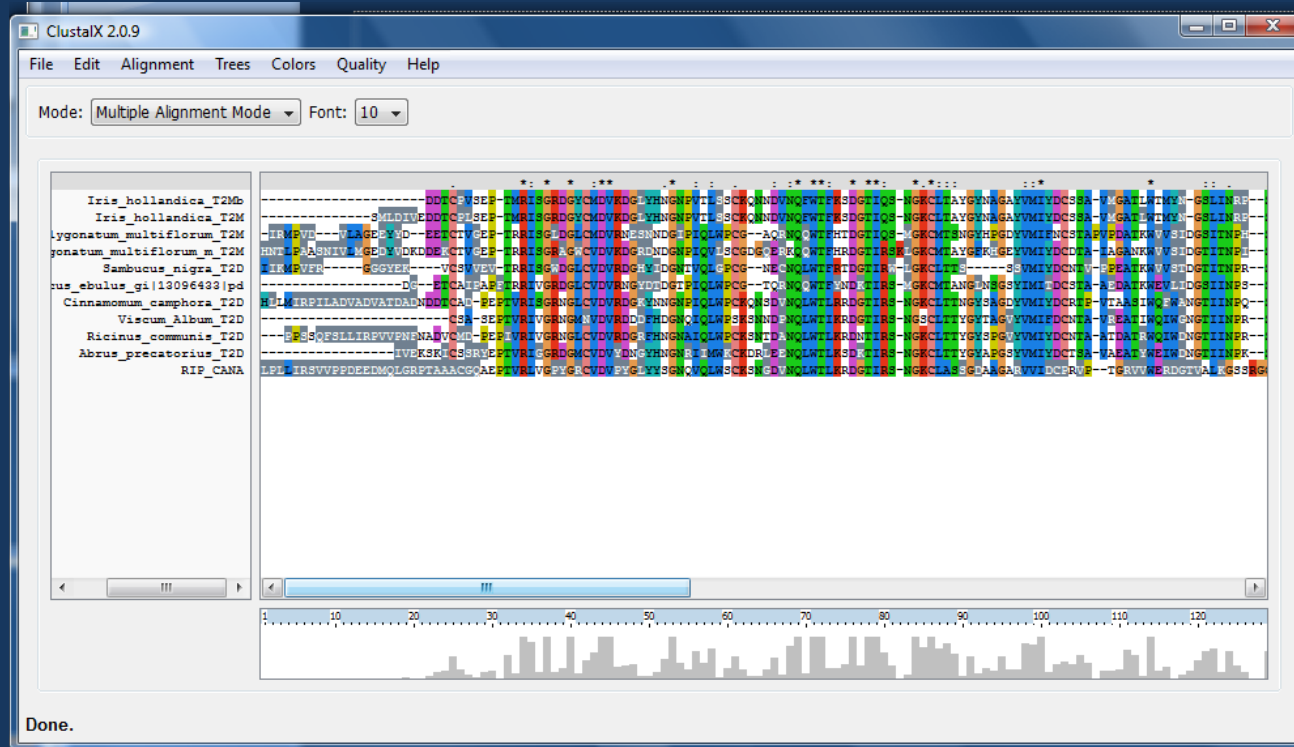


Símbolos indicam conservação de resíduos ou grupos de resíduos em uma coluna

"*" indicates positions which have a single, fully conserved residue.
 ":" indicates that one of the following 'strong' groups is fully conserved:
 STA
 NEQK
 NHQK
 NDEQ
 QHRK
 MILV
 MILF
 HY
 FYW
 ". " indicates that one of the following 'weaker' groups is fully conserved:
 CSA
 AIV
 SAG
 STNK
 STPA
 SGND
 SNDEQK
 NDEQHK
 NEQHRK
 FVLIM
 HFY

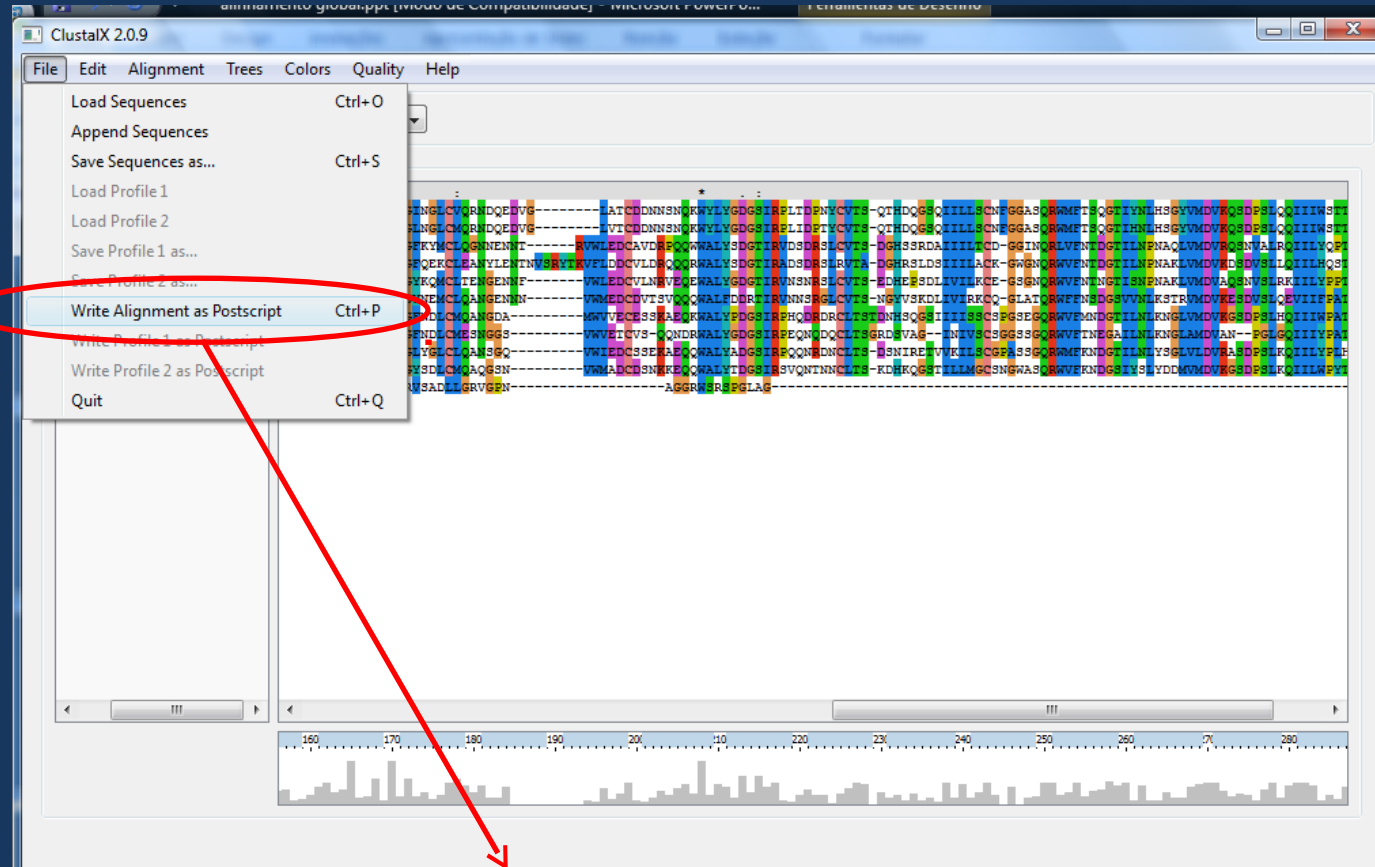
Ok

Resultado de alinhamento ClustalX



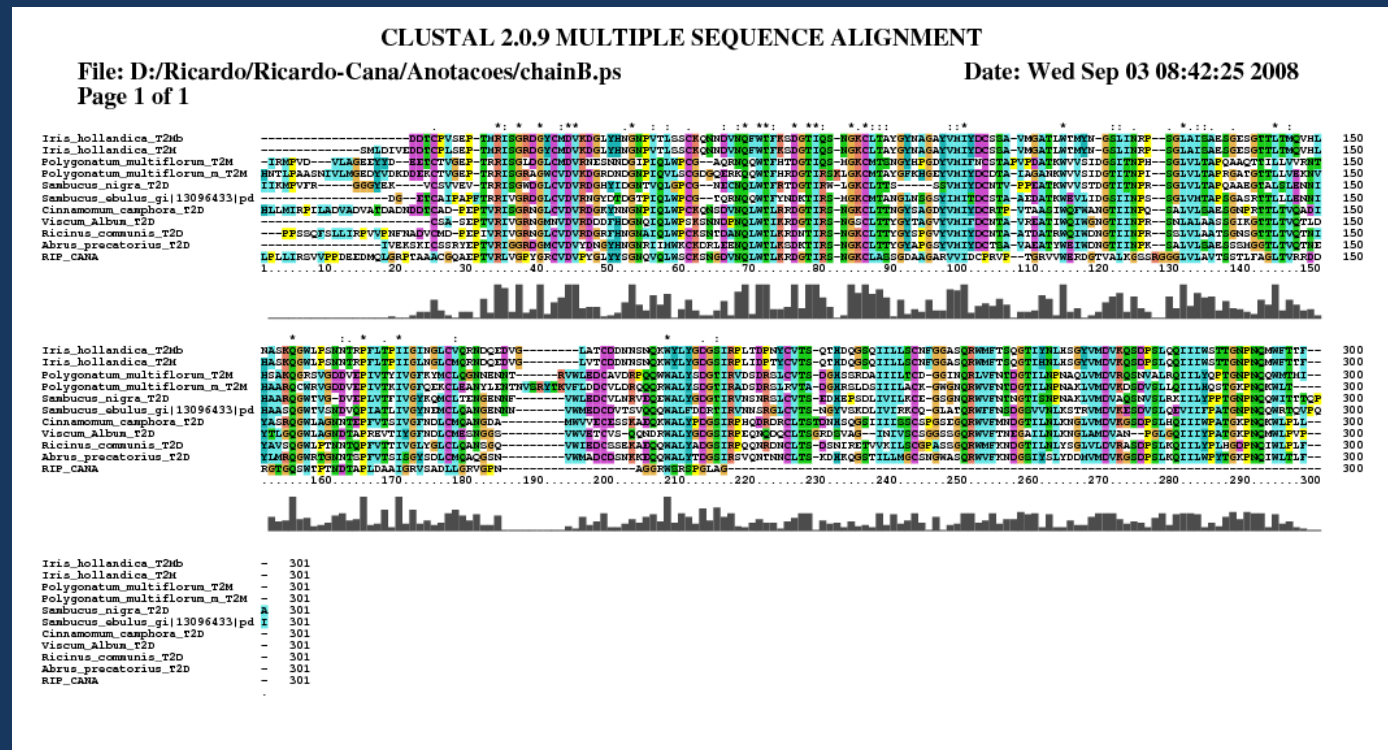
- Low scoring segments- existe a opção de mostrar em cinza regiões com baixo score e que portanto não seriam muito confiáveis

Visualização do resultado de alinhamento ClustalX



Exporta o alinhamento em formato postscript

Visualização do resultado de alinhamento ClustalX



- Resultado de alinhamento em arquivo postscript

Visualização do resultado de alinhamento ClustalX

The screenshot shows the BOXSHADE 3.21 web interface. At the top is a navigation bar with links: Home, Services, Courses, Links, and Contacts. The title is 'BOXSHADE 3.21' with the subtitle 'Pretty Printing and Shading of Multiple-Alignment files'. A note mentions a server update on July 17. Below the note is a list of supported output formats: Postscript/EPS, RTF old/new, XFIG, ASCII, HPGL, and PICT. The interface includes several input fields: 'Output format' (set to RTF_new), 'Font Size' (10), 'Consensus Line' (no consensus line), 'Fraction of sequences' (0.5), 'Enter sequence numbers' (empty), and 'Query title (optional)' (empty). A warning message states: 'When pasting MSF or ClustalW files, please make sure that the pasted text starts with the header line of the alignment and contains no extra blank lines at the bottom.' The 'Input sequence format' is set to ALN. The 'Paste your multiple-alignment file' section shows a text area with 'RIP_CANA' and a file selection button. Below the text area, the file 'Iris_hollandica_T2Mb' is selected. At the bottom are 'Run BOXSHADE' and 'Clear Input' buttons.

ch.EMBnet.org
Home Services Courses Links Contacts

BOXSHADE 3.21

Pretty Printing and Shading of Multiple-Alignment files

Notes: Starting July 17, version 3.21 is running on this server. The only changes are some improvements in the RTF output routine that enables RTF with shaded background for users of MS-Word 7.0.
I have also compiled a list of [list of frequently asked questions](#).
This server uses version 3.21 of BOXSHADE, written by K. Hofmann and M. Baron. [BOXSHADE](#) is in the public domain and available from Source Forge <http://sourceforge.net/projects/boxshade/>. The available version runs on PCs, VMS- and OSF1-machines and includes much more options than are implemented on this server.
This **server** takes a multiple-alignment file in either GCG's MSF-format or Clustal's ALN-format. Output can be created in the following formats:

- Postscript/EPS (using shaded background)
- RTF old (using colors)
- RTF new (using shaded background)
- XFIG-files (using shaded background)
- ASCII (showing similarities)
- ASCII (showing differences)
- HPGL (using colors)
- PICT (for later editing on MACs and PCs)

If you have problems using this server (like getting no result), [read this](#) and see the [FAQ list](#).

Output format: **RTF_new**
Font Size: **10**
Consensus Line: **no consensus line**
Fraction of sequences: **0.5** (that must agree for shading)
Enter sequence numbers:
Query title (optional):
When pasting MSF or ClustalW files, please make sure that the pasted text starts with the header line of the alignment and contains no extra blank lines at the bottom.

Input sequence format: **ALN**
Paste your multiple-alignment file (see above for valid formats):
RIP_CANA
Iris_hollandica_T2Mb
!!!

Box shade- ferramenta para visualização de alinhamentos. Aceita formato aln (clustal). Pode exportar em formato rtf

http://www.ch.embnet.org/software/BOX_form.html

Visualização do resultado de alinhamento ClustalX

```
Iris_hollandica_T2Mb.....1-----DITCPVSEP--TMRISGRDGYCMVVDKGLVHN--GNPVTSS-  
Iris_hollandica_T2M.....1-----SMLDIVEDDITCPVSEP--TMRISGRDGYCMVVDKGLVHN--GNPVTSS-  
Polygonatum_multiflorum_T2M.....1--IRMPVD---VLAGEEYYD--EETCTVGEF--TMRISGLDGLCMVVRNENNDGTPVQLWP-  
Polygonatum_multiflorum_m_T2M.....1--HNTLPAAASNVLMGEDYVDKDDKCTVGEF--TMRISGRDGYCMVVDKGLVHN--GNPVTSS-  
Sambucus_nigra_T2D.....1--IHKMPVFR-----GGGYEK-----VCSVVEV--TMRISGRDGLCVDVVRDGHVVDGATVQLWP-  
Sambucus_ebulus_gi|13096433|pd.....1-----DG--ETCAHPAPFTRRI--TMRISGRDGLCVDVVRNENNDGTPVQLWP-  
Cinnamomum_camphora_T2D.....1--HLLMIRPILADVADVATDADND--TCAD--PEPTVRI--TMRISGRDGLCVDVVRDGHVVDGATVQLWP-  
Viscum_Album_T2D.....1-----CSA--SEPTVRI--TMRISGRDGLCVDVVRDGHVVDGATVQLWP-  
Ricinus_communis_T2D.....1-----PPSSQFSLIRPVVNFMA--VCM--PEPTVRI--TMRISGRDGLCVDVVRDGHVVDGATVQLWP-  
Abrus_precatorius_T2D.....1-----IVEKSKTCSRYEPTVRI--TMRISGRDGLCVDVVRDGHVVDGATVQLWP-  
RIP_CANA.....1--LPLLIRSVVPPDEEDMQLGRPTAAACQAEPTVRI--TMRISGRDGLCVDVVRDGHVVDGATVQLWP-  
  
Iris_hollandica_T2Mb.....39--CKQNNNDVNCQFWIFRSOGTITCS--NGKCLTAYGYNAGAYVMIYDCSSA--VMGATILATMNN--E-  
Iris_hollandica_T2M.....46--CKQNNNDVNCQFWIFRSOGTITCS--NGKCLTAYGYNAGAYVMIYDCSSA--VMGATILATMNN--E-  
Polygonatum_multiflorum_T2M.....54--CG--AQRNQCWTFPHDGTITCS--MGKCHTANGYHFGDYVMIN--CSTAPUPDATKAVUSIIC-  
Polygonatum_multiflorum_m_T2M.....60--CGDGERKQWTFPHDGTITCS--MGKCHTANGYHFGDYVMIN--CSTAPUPDATKAVUSIIC-  
Sambucus_nigra_T2D.....51--CG--NEONQLWTFPHDGTITCS--MGKCHTANGYHFGDYVMIN--CSTAPUPDATKAVUSIIC-  
Sambucus_ebulus_gi|13096433|pd.....41--CG--TORNQCWTFPHDGTITCS--MGKCHTANGYHFGDYVMIN--CSTAPUPDATKAVUSIIC-  
Cinnamomum_camphora_T2D.....60--CKQNNNDVNCQFWIFRSOGTITCS--NGKCLTAYGYNAGAYVMIYDCSSA--VMGATILATMNN--E-  
Viscum_Album_T2D.....36--SKSNNDVNCQFWIFRSOGTITCS--NGKCLTAYGYNAGAYVMIYDCSSA--VMGATILATMNN--E-  
Ricinus_communis_T2D.....57--CKSNNDVNCQFWIFRSOGTITCS--NGKCLTAYGYNAGAYVMIYDCSSA--VMGATILATMNN--E-  
Abrus_precatorius_T2D.....44--CKDRLEDNQLWTLRDTITCS--NGKCLTAYGYNAGAYVMIYDCSSA--VMGATILATMNN--E-  
RIP_CANA.....61--CKSNNDVNCQFWIFRSOGTITCS--NGKCLTAYGYNAGAYVMIYDCSSA--VMGATILATMNN--E-  
  
Iris_hollandica_T2Mb.....96--SLINRP--SGIAISABSGES--ETILTQVH--HASKQGLPSSNTRPFLTPITIGLGLCVR-  
Iris_hollandica_T2M.....103--SLINRP--SGIAISABSGES--ETILTQVH--HASKQGLPSSNTRPFLTPITIGLGLCVR-  
Polygonatum_multiflorum_T2M.....111--SITNPH--SGIWLAPAPQAQITILTQVH--HASKQGLPSSNTRPFLTPITIGLGLCVR-  
Polygonatum_multiflorum_m_T2M.....119--TIINPR--SGIWLAPAPQAQITILTQVH--HASKQGLPSSNTRPFLTPITIGLGLCVR-  
Sambucus_nigra_T2D.....102--TIINPR--SGIWLAPAPQAQITILTQVH--HASKQGLPSSNTRPFLTPITIGLGLCVR-  
Sambucus_ebulus_gi|13096433|pd.....97--SIINPS--SGIWLAPAPQAQITILTQVH--HASKQGLPSSNTRPFLTPITIGLGLCVR-  
Cinnamomum_camphora_T2D.....118--TIINPR--SGIWLAPAPQAQITILTQVH--HASKQGLPSSNTRPFLTPITIGLGLCVR-  
Viscum_Album_T2D.....94--TIINPR--SGIWLAPAPQAQITILTQVH--HASKQGLPSSNTRPFLTPITIGLGLCVR-  
Ricinus_communis_T2D.....115--TIINPR--SGIWLAPAPQAQITILTQVH--HASKQGLPSSNTRPFLTPITIGLGLCVR-  
Abrus_precatorius_T2D.....102--TIINPR--SGIWLAPAPQAQITILTQVH--HASKQGLPSSNTRPFLTPITIGLGLCVR-  
RIP_CANA.....118--TIINPR--SGIWLAPAPQAQITILTQVH--HASKQGLPSSNTRPFLTPITIGLGLCVR-
```