Introduction

Multiple sequence alignment is one of the most commonly employed computational tools in biology. It has been used extensively to demonstrate sequence homologies between structurally and functionally related proteins and to aid in the determination of evolutionary relationships between proteins within and between species. In general, proteins which show a high degree of functional conservation in the course of evolution also show a high degree of sequence similarity and can therefore be aligned with a good degree of biological accuracy. However, proteins that have retained small regions of high similarity but have otherwise undergone extensive modification in intervening regions may not be aligned in a biologically accurate manner by the most of the commonly used alignment algorithms[1]. This presents a particularly great challenge if the intervening sequences are composed of natively disordered regions for which there is little selective pressure to maintain a given domain structure. This is the case for the Cubitus interruptus/Gli family of zinc-finger transcription factors, which are involved in cellular responses to the Hedgehog morphogen.

The Cubitus interruptus (Ci)/Gli family of zinc-finger transcription factors is widely conserved both in form and in general function throughout the major metazoan taxa and is involved in the numerous Hedgehog-dependent developmental processes. In Drosophila sp., Ciona sp. and Amphioxus sp, this family is represented by one protein, Cubitus interruptus in fly, CionaGli in Ciona and AmphiGli in amphioxus. In vertebrates, the family is represented by at least 3 members, Gli1, 2 and 3, which are thought to have arisen as a result of two successive rounds of gene duplication, much in the same way as the developmentally important Hox gene clusters. From sequence comparison and phylogenetic reconstruction, it appears that the earliest duplication separated Gli1 from Gli2/3, the second separating Gli2 and Gli3. The vertebrate Gli proteins may be thought of as dividing between themselves the tasks of the ancestral single Ci/Gli protein. For instance, all Ci/Gli protein products contain a highly conserved (90% identity) zinc-finger DNA binding domain, a conserved binding motif for Suppressor of Fused and several PKA phosphorylation motifs spaced at irregular intervals in the C-terminus. In addition, they have each evolved novel functions and play individually important and distinct roles in the development of the vertebrate limb, central nervous system and other systems. Understanding which residues and sequence motifs of the ancestral Ci/Gli protein have been conserved in each of the vertebrate paralogs is a key step in predicting a mechanistic basis for novel evolved functions.

As described above, the Ci/Gli proteins contain of a highly conserved set of 4 zinc-finger motifs, which directly bind to conserved Hedgehog response elements, and a fifth zinc finger not directly involved in DNA binding. The zinc fingers are flanked by N- and C-terminal regions that are thought to bind accessory proteins involved in modulating the activity, location and stability of the Ci/Gli proteins in response to Hedgehog signaling. Genetic studies in Drosophila have revealed a number of proteins that potentially interact with Ci/Gli but the nature of these interactions is poorly understood in a majority of cases. One exception is the interaction of Ci/Gli with the Suppressor of Fused (SuFu) protein. SuFu is thought to recognize a highly conserved pentapeptide motif (SYGHL/I) in the region N-terminal to the zinc-finger repeats. Much of the intervening sequence is poorly conserved, especially between paralogs but even between orthologs in different vertebrate classes. In addition, there are several conserved regions in the C-terminus including a variable number of conserved PKA phosphorylation sites, either independent from or in association with previously uncharacterized conserved sequences. Adjacent to the zinc-finger domains in the N-terminus is a conserved motif (D-S-G-V/m-E/d-M/v-XXN) of unknown function that appears to have arisen and been preserved in the chordate lineage, including two ancestral chordate Gli proteins and all examined Gli1 and 2 proteins. Much of the remaining portions of the Ci/Gli proteins lack clear regions of sequence conservation. It is unknown what function, if any, these other sequences play in the conserved, let alone the unique functions of these proteins. It is tempting to write them off as inconsequential "space fillers", but in the absence of a detailed understanding of the spacing requirements of conserved motifs, this would be a foolish approach. If multiple sequence analysis is to provide any clues to the function of these non-conserved regions, it is critical that alignments be generated that consistently align the conserved regions. In this paper I will examine the ability of commonly used alignment algorithms to properly (and in the same alignment) align these three classes of conserved motifs in a set of Ci/Gli sequences from representative metazoan taxa. I will demonstrate that the most commonly used algorithms, including ClustalW and T-Coffee fail to consistently produce proper alignments. I will examine the causes underlying this failure and identify alternate algorithms and modifications to common algorithms which can overcome these underlying weaknesses.

Methods

Sequences

Ci/Gli sequences were chosen from two metazoan phyla: Arthropoda and Chordata. 1 class of Arthropoda (Insecta) was represented by Drosophila melanogaster. Three subphyla of Chordata were represented by Ciona intestinalis (Urochordata), Branchiostoma floridae (aka Amphioxus, Cephalochordata) and Craniata. Three classes of Craniata were represented by Danio rerio (ray-finned fishes, Actinopterygii), Xenopus laevis (Amphibia) and Homo sapiens (Mammalia). Drosophila, Ciona and Amphioxus each possess one known member of the Ci/Gli family. These were represented by the following sequences: NP_524617.2 (Cubitus interruptus [Drosophila melanogaster]), CAB96572.1 (AmphiGli protein [Branchiostoma floridae]). The available sequence for Drosophila Ci is likely to represent the biologically complete sequence whereas the only available AmphiGli sequence appears to lack approximately 200 N-terminal residues judging from the available complete Arthropod and Chordate homologs. The sequence for Ciona intestinalis was obtained by a tblastn (protein vs. translated nucleotide blast) query (with the Branchiostoma floridae AmphiGli peptide sequence) of the DOE Joint Genome Institute Version 1.0 release of the Ciona intestinalis genome. The resultant putative exons were assembled into a virtual mRNA and translated with standard sequence analysis software. The resulting peptide sequence is 641 amino acids in length and encompasses the SYGHL pentapeptide and the five zinc fingers as well as a considerable portion of the C-terminus. Approximately 600 amino acids comprising putative N-terminal and extreme C-terminal residues are likely to be missing. For Xenopus the sequences used were as follows: Q91690 (Gli1), AAD28180 (Gli2), Q91660 (Gli3). For Zebrafish the sequences used are as follows: AAO43495 (Gli1, Detour), NP_571042 (Gli2, you-too). A Zebrafish homolog of Gli3 has not yet been identified. For Human, the sequences used are as follows: P08151 (Gli1), NP_000159 (Gli3). All of the available human Gli2 sequences lack a large portion of the N-terminus (including the SYGHL pentapeptide) when compared to amphibian and fish homologs and were therefore not used for this analysis.

Alignment tools and algorithms

Pairwise alignments were performed with the Gap and Best Fit tools on the GCG SeqWeb website. These tools use the Needleman-Wunsch (global) and Smith-Waterman (local) alignment algorithms respectively. In each case the BLOSUM62 scoring matrix was used with a Gap Opening Penalty (GOP) of 8 and a Gap Extension Penalty (GEP) of 2. End gaps were not penalized. Pairwise alignments were also performed using the Pairwise BLAST [2]function on the NCBI website. Again, the BLOSUM62 scoring matrix was used with a GOP of 11 and a GEP of 1, a word size of 3 and an Expectation of 10. Multiple alignment was performed in 5 different ways, 3 of which represented pair-wise progressive algorithms (Pileup, 2 conditions of ClustalW[3, 4]), one a consistency based progressive algorithm (T-Coffee[5]) and one that employed a segment-based progressive approach (DiAlign[6]). For alignments with Pileup (used on the GCG SeqWeb site) scoring matrix and gap settings were identical to those used for Best Fit and Gap. ClustalW alignments were performed on the DeCypher server with the BLOSUM62, BLOSUM85 and BLOSUM100 matrices, Ktuple size set at 1, window size at 5, pairwise gap penalty at 3, GOP at 10, GEP at 0.05, residue specific gaps ON, hydrophilic gaps ON, gap separation distance of 8, NO endgap penalty. ClustalW alignments were run twice, either WITH or WITHOUT negative matrix values. The DiAlign alignments were performed on the Genomatix server with a threshold of 0.00. None of the sequences were edited or modified prior to alignment except where otherwise stated.

Results

Alignments were performed on four Ci/Gli sequences (AmphiGli, CionaGli, Drosophila Ci, Zebrafish Gli1) to determine the ability of the different multiple sequence alignment algorithms to correctly align the SuFu pentapeptide. The sequences used were chosen to represent four major taxonomic groups, so that there would be as little overweighting bias as possible from phylogenetically related sequences. Resulting alignments are shown in Figure 1. The SuFu pentapeptide is highlighted in orange. An eight amino acid sequence {(R/K)KR(A/P)LS(I/S)S) N-terminal to the SuFu pentapeptide motif was selected as an alignment reference based on its conservation in three of the sequences. Its conservation in AmphiGli could not be determined due to the fact it lay N-terminal to the available AmphiGli sequence. Of the five algorithms tested, only ClustalW (with negative matrix values turned off, ClustalW^{off}) was *unable* to properly align the SuFu motif (Fig 1A). ClustalW^{off} was furthermore unable to produce a single pairwise alignment of the motif *within* the multiple sequence alignment. When individual ClustalW^{off} pairwise alignments were performed (data not shown) 3 out of the 6 possible pairwise alignments did not show an alignment of the SuFu motif. The 8aa upstream motif was likewise improperly aligned. Engaging the negative matrix values in ClustalW improved this algorithm's alignment performance for the SuFu motif. However, the 8aa upstream motif was left unaligned. Like ClustalW, Pileup is a progressive alignment algorithm which assembles a multiple alignment from individual pairwise alignments in the sequence set. Despite the similarity of the algorithms Pileup (Fig. 1D) perfectly aligned both motifs in all relevant sequences. The same held true for both T-Coffee and DiAlign, which gave perfect alignments.

ClustalW^{off} is in this instance insensitive to consecutive identical residues. This insensitivity appears to have arisen at the stage of the initial pairwise alignments and was carried over into the assembly of the multiple alignment from these pairwise analyses. This conclusion is supported by the observation that removal of the AmphiGli sequence, which itself was involved in 2 of the 3 pairwise alignment failures, allowed ClustalW to correctly align the SuFu motif. T-Coffee, though it is based on ClustalW pairwise alignments succeeded in producing correct alignments, probably due to its consistency approach, which avoids early commitment to false pairwise alignments. Because each of the sequences was involved in at least one correct pairwise alignment with another sequence, T-Coffee was able to correctly assemble the multiple alignment. DiAlign uses a segment-based approach, which instead of comparing individual residues in pairwise alignments, searches for discrete regions of similarity and builds an alignment from non-overlapping segment pairs. The authors of the DiAlign algorithm recommend its use in instances of sequences containing islands of conserved residues in a background of low overall sequence similarity. This is exactly the situation observed with Ci/Gli proteins and it is therefore not surprising that DiAlign would be an effective tool for producing accurate alignments with these proteins.

Several different combinations of sequences were used in order to assess the sequence dependence of each of the algorithms in aligning the SuFu motif. The results are summarized in Table 1. Different combinations of sequences differentially effect each of the algorithms, with the AmphiGli sequence associated with a majority of the alignment failures. This suggests that the AmphiGli sequence may either be too divergent from the other sequences used in the alignments or it may have other peculiarities which force false alignments. It seems, however, that much of the alignment difficulty arises from the fact that the available AmphiGli sequence lacks much of the putative N-terminus. This is illustrated by the fact that editing each of the sequences to remove N-terminal residues up to the SuFu motif leads to a perfect alignment of the motif with all evaluated algorithms (data not shown). This suggests that there is sufficient global amino acid similarity in the N-terminal region of the other sequences to create “decoy” diagonals in pairwise alignments, thus leading to the assembly of incorrect multiple alignments of the SuFu motif. This is illustrated in Figure 2, which graphically compares the diagonals produced with three different scoring matrix stringencies. Using the BLOSUM62 and even BLOSUM85 scoring matrix, there are numerous competing diagonals. Most of these are off of the main diagonal (as defined by the zinc finger consensus region) and in the absence of a penalty for end gaps they should be ignored in the alignment assembly. Nonetheless, it is clear that ClustalW (Figure1) is unable to choose the correct diagonal for the SuFu motif. Increasing the scoring matrix stringency to BLOSUM100 drastically reduces the number of diagonals, suggesting that using an identity matrix in the ClustalW alignment would produce a more biologically accurate result. This is not the case however as illustrated by Figure 3. Therefore it appears that invoking negative matrix values in ClustalW alignments or using a segment-based algorithm such as DiAlign is crucial to producing an alignment of highly conserved short motifs in a background of high amino acid similarity.

However, these other approaches are not without shortcoming when challenged by very short conserved motifs in a background of similar amino acids. This is illustrated by attempts to align the PKA phosphorylation motif region in the C-terminal region of the Ci/Gli proteins (Figure 4). The PKA phosphorylation consensus sequence is a short tetrapeptide motif (R-R/k-X-S). Assuming a 5% frequency for R and K and a frequency of 8% for S, this motif can be expected to occur once by chance in a sequence of 2000 amino acids. In the four sequences shown in Figure 4, this motif appears between 5 and 7 times in an approximately 200 amino acid region, or approximately 50 to 70 times the frequency one would expect by chance. This observation combined with experimental evidence suggesting an important role for PKA in controlling the stability of the Ci/Gli proteins argues strongly that these motifs are biologically relevant and likely to represent evolutionarily conserved positions in the protein. Unfortunately, none of the alignment algorithms shows much success in aligning all these motifs. The consistency-based approach of T-Coffee (Fig. 4C) seems to be more efficient than ClustalW (either with (Fig 4B) or without (Fig. 4A) negative matrix values). DiAlign (Fig 4D) and Pileup (Fig 4E) are both more effective than ClustalW but show curious alignment errors (of one or two residues) in otherwise very closely aligned sequences. This problem is well illustrated by re-examining the graphical alignment in Figure 2 which shows a high number of competing diagonals in this region, even using a high stringency scoring matrix such as BLOSUM100. On the other hand, a motif of unknown function(D-S-G-V/m-E/d-M/v-XXN), found in the C-terminus of Gli proteins (only in the chordate lineage) is properly aligned by all of the alignment algorithms except Pileup. This performance is notable in light of the motif's incomplete sequence conservation, variable distance from the nearest conserved sequences in the Zinc finger domain and its absence from Drosophila Ci. It appears that motif length may be playing a crucial role in assuring proper alignment in this region of the protein. This conclusion is further supported by the observation that another conserved motif of unknown function (F/SYDPIS) is properly aligned despite its complete absence from the Zebrafish Gli1. (This motif is conserved in all observed Ci/Gli proteins with the exception of all known Gli1 proteins.)

Discussion

Most of the multiple sequence alignment algorithms in use today are extremely efficient at correctly aligning structurally similar yet distantly related proteins such as vertebrate myoglobin and plant leghemoglobin. Much of the power of these algorithms is derived from the use of "biologically correct" amino acid substitution scoring matrices such as BLOSUM. These matrices allow for comparisons based on structural/functional similarities between amino acids and are excellent at producing alignments between distantly related proteins when there are evolutionary constraints on the structural/functional character of a given protein or protein region. Put another way, there are a limited number of ways to construct a globin, and because of this, alignments between them can be reproducibly derived with currently available tools. There are, however, almost infinite ways to construct a natively disordered random coil and almost as many ways to align such sequences. This seems to be the case with the Ci/Gli class of transcription factors despite a significant number of distinct and biologically important conserved motifs. I have demonstrated that four of the currently available multiple alignment algorithms are capable of aligning unique conserved sequences of five amino acids or greater. All of them fail however when attempting to align repeated, tripeptide motifs such as the PKA phosphorylation site. It is clear that modifications to the alignment algorithms are necessary in order to perform *ab initio* alignments of these and similar sequences.

In the course of this study, it was noticed that the BLAST algorithm consistently produced good alignments of the PKA motifs when querying database sequences. Based on these observations, a series of pairwise alignments was carried out between four of the Ci/Gli sequences using the pairwise BLAST tool on the NCBI website. The results from these alignments are shown in Figure 5. With the exception of two instances, those PKA motifs which are aligned, are aligned properly. When compared to pairwise alignments produced with any of the other algorithms, Pairwise BLAST was far superior in aligning the PKA motifs. This result suggested a possible modification that could be made to an otherwise superior algorithm such as T-Coffee. T-Coffee is similar to ClustalW and PileUp in that it produces a multiple sequence alignment by progressive pairwise alignments. ClustalW aligns sequences in an order based on a their relatedness to the other sequences as determined by construction of a sequence similarity "guide tree" [3, 4]. Once a sequence is aligned by ClustalW it cannot be unaligned even if its alignment conflicts with that of subsequent sequences. T-Coffee overcomes this defect by checking each alignment for consistency against a "library" of ClustalW global alignments and Lalign local alignments of the sequences. The basic structure of the algorithm is shown in figure 6A[5].

There are several steps in this algorithm that are potential targets for improvement. Following the production of the global and local libraries, T-Coffee compares their results and determines the number of times given pairs of residues are aligned in the libraries and assigns weights based on the frequency of pairing. This step is necessarily susceptible to systematic alignment biases in the algorithms used to construct the reference libraries. If the algorithms are unable to detect and align conserved motifs, the weights will reflect this and the motifs will not be aligned properly in the final output. This bias in the case of short and rare sequence motifs could be dealt with, as described above, through the use of pairwise BLAST in the construction of the library of local alignments. Furthermore, endowing the algorithm with the ability to recognize important alignments, or to give greater weights to alignments of known motifs, would vastly improve the probability of aligning short and rare motifs. This would not require that the program know anything about the specific sequences being aligned. The algorithm could take advantage of a curated library of known sequence motifs and variants. As a first step in the algorithm the sequences could be queried for the presence of known motifs and a sequence set-specific motif library could be produced. When global and local alignments are compared in subsequent steps in order to assign residue-specific weights, the algorithm could refer to the sequence specific library and reward exact motif matches with a high weight.

Such an algorithm would rely on the appearance of motif alignments in at least one of the reference library sequence pairs. There is always the chance that these alignments will not occur as a result of weaknesses in the algorithms used to construct the libraries or anomalies in the sequences chosen for the alignments. Further modifications to the algorithm could be made to compensate for these shortcomings. For instance the algorithm could query a sequence database with the BLAST algorithm to determine similar sequences not represented in the user provided sequence set. These query derived sequences could be used to construct the pairwise global and local libraries, but would not appear in the final alignment. Ideally, multiple alignment algorithms would be able to take advantage of sequence annotation in order to properly constrain alignments around conserved sequences. This, of course, relies on the existence of sequence annotation and would be less useful for novel sequences and poorly understood families. Nonetheless, given the ever increasing amount of sequence annotation, its application to sequence alignment algorithms would be potentially very useful. Developing methods to integrate this sort of data into sequence analysis is the next challenge in multiple sequence alignment.

References

1. **Notredame, C.,** ***Recent progress in multiple sequence alignment: a survey.*** **Pharmacogenomics, 2002. 3(1): p. 131-44.**
2. **Altschul, S.F., et al.,** ***Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.*** **Nucleic Acids Res, 1997. 25(17): p. 3389-402.**
3. **Thompson, J.D., D.G. Higgins, and T.J. Gibson,** ***Improved sensitivity of profile searches through the use of sequence weights and gap excision.*** **Comput Appl Biosci, 1994. 10(1): p. 19-29.**
4. **Thompson, J.D., D.G. Higgins, and T.J. Gibson,** ***CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.*** **Nucleic Acids Res, 1994. 22(22): p. 4673-80.**
5. **Notredame, C., D.G. Higgins, and J. Heringa,** ***T-Coffee: A novel method for fast and accurate multiple sequence alignment.*** **J Mol Biol, 2000. 302(1): p. 205-17.**
6. **Morgenstern, B., et al.,** ***DIALIGN: finding local similarities by multiple sequence alignment.*** **Bioinformatics, 1998. 14(3): p. 290-4.**

Table 1: **Alignment of SuFu pentapeptide**

	Pile-up	ClustalW(Neg OFF)	ClustalW (Neg ON)	T-Coffee	DiAlign
AmphiGli	1	0	1	1	1
CionaGli	1	0	1	1	1
DmCi	1	1	1	1	1
DrGli1	1	1	1	1	1
HsGli1	1	1	1	1	1
XlGli1	1	1	1	1	1
AmphiGli	0	0	1	0	1
DmCi	1	1	1	1	1
DrGli1	1	1	1	1	1
HsGli1	1	1	1	1	1
XlGli1	1	1	1	1	1
AmphiGli	1	0	1	1	1
CionaGli	1	0	1	1	1
DmCi	1	0	1	1	1
DrGli1	1	0	1	1	1
AmphiGli	0	1	1	0	1
DmCi	0	1	1	1	1
DrGli1	0	0	1	1	1
Ciona	1	1	1	1	1
DmCi	1	1	1	1	1
DrGli1	1	1	1	1	1
AmphiGli	1	0	1	0	1
DmCi	1	1	1	1	1
DrGli2	1	1	1	1	1
XlGli2	1	1	1	1	1
AmphiGli	1	0	1	0	1
DmCi	1	0	1	1	1
HsGli3	1	1	1	1	1
XlGli3	1	1	1	1	1
Correct Alignments	25	18	29	25	29

Dr = Danio r. (Zebrafish)
Dm = Drosophila m.
Hs = Homo sapiens
Xl = Xenopus l.

Figure 1: Alignment of N-terminal region of 4 Ci/Gli proteins

SYGHL/I colored in orange for visibility. (R/K)KR(A/P)LS(I/S)S colored in pink.

A. ClustalW alignment, negative matrix values OFF

```
AmphiGli        ----------------------------------------ASTGSYGHLSASAMRTESGA 20
CionaGli        -----------------------ARKRPLSISPCFSDTGLDITAMIRTSPNSLLPFGGIA 37
DmCi            FHFSVDGNRRLGSPRPPGGSIRASISRKRALSSSPYSDSFDINSMIRFSPNSLATIMNGS 240
DrGli1          PPHSMMGHRGMPPPEGMSGAPYCNQNMMTSHHNLPHNQHTSELMASGDASCFSTPRSMLK 136

AmphiGli        ESKPGDPVLRKHAVQRADAHVPVPTSP--------------------------------- 47
CionaGli        NSRSSSVASGGSYGHLAAGGISSIFSSK-------------------------------- 65
DmCi            RGSSAASGSYGHISATALNPMSHVHSTRLQQIQAHLLRASAGLLNPMTPQQVAASGFSIG 300
DrGli1          LSKKRALSISPLSDASVDLQTVIRTSPN-------------------------------- 164

AmphiGli        -------AMQQFHNRLMRQKSPFHFGMPHASPFAAPLPAGMAM----------------- 83
CionaGli        -------PTYKVVLYTFILPNALYFSSPTFGYQTPMMTSPQHL-------H-------AH 104
DmCi            HMPTSASLRVNDVHPNLSDSHIQITTSPTVTKDVSQVPAAAFSLKNLDDAREKKGPFKDV 360
DrGli1          -------SLVAFVNSRCGPNNPSSYGHLSVGTMSPSLGFSSSINYSRPQGNIYSHPVPSC 217
```

B. ClustalW alignment, negative matrix values ON

```
AmphiGli        ------------------------------------------------------------
CionaGli        ----ARKRPLSISPCFSDTGL-------------------DITAMIRTSPNSLLPFGGIA 37
DmCi            FHFSVDGNRRLGSPRPPGGSIRASISRKRALSSSPYSDSFDINSMIRFSPNSLATI---M 237
DrGli1          SMLKLSKKRALSISPLSDASV-------------------DLQTVIRTSPNSLVAF---- 169

AmphiGli        -------ASTGSYGHLSASAMRTESGAESKPGDPVLRKHAVQRADAHVPV---------- 43
CionaGli        NSRSSSVASGGSYGHLAAGGISSIFSSKP------------T------------------ 67
DmCi            NGSRGSSAASGSYGHISATALNPMSHVHSTRLQQIQAHLLRASAGLLNPMTPQQVAASGF 297
DrGli1          VNSRCGPNNPSSYGHLSVGTMSPSLGFSS------------------------------- 198
```

C. T-Coffee alignment

```
AmphiGli      .......... .......... .......... .......... ..........
CionaGli      .......... .......... ARKRPLSISP CFSDTGLDIT AMIRTSPNSL
DmCi          DGNRRLGSPR PPGGSIRASI SRKRALSSSP .YSD.SFDIN SMIRFSPNSL
DrGli1        ........PR S.....MLKL SKKRALSISP .LSDASVDLQ TVIRTSPNSL

AmphiGli      .......... ....ASTGSY GHLSASAMRT ESGAESK... ..PGDPVLRK
CionaGli      LPFGGIANSR SSSVASGGSY GHLAAGGISS IFSSKPT..Y ...KVVLYTF
DmCi          ATI...MNGS RGSSAASGSY GHISATALNP MSHVHST..R ...LQQIQAH
DrGli1        VAF...VNSR CGPNNPS.SY GHLSVGTMSP SLGFSSSINY SRPQGNIYSH
```

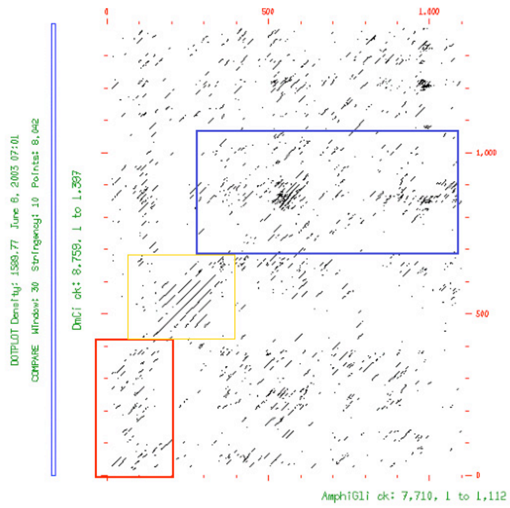
D. DiAlign alignment

```
AmphiGli      1   ---------- ---------- ---------- ---------- ----------
CionaGli      1   ---------- ---------- ---------- ---ARKRPLS ISPcfSDTGL
DmCi        173   agslastdfh fsvdgnrrlg sprppggsir asISRKRALS SSPY-SD-SF
DrGli1      137   ---------- ---------- ---------- --LSKKRALS ISPL-SDASV

AmphiGli      1   ---------- ---------- -------AST GSYGHLSASA MRTESGAESK
CionaGli     18   DITAMIRTSP NSLLPFggIA NSRSSSVASG GSYGHLAAGG ISSIFSSKPT
DmCi        221   DINSMIRFSP NSLat---IM NGSRGSSAAS GSYGHISATA LNPMSHVHST
DrGli1      154   DLQTVIRTSP NSLVAFvnsr cgpn----NP SSYGHLSVGT MSPSLGFSSS
```

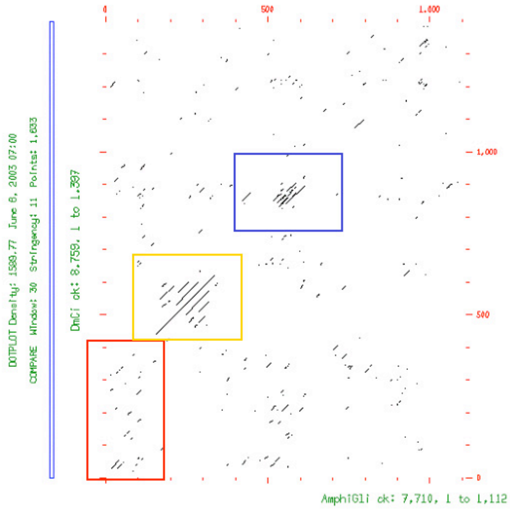
E. Pileup alignment

```
             201                                                250
CIONAGLI     ~~~~~~~~~~ ~~~ARKRPLS ISPCFSDTGL DITAMIRTSP NSLLPFGGIA
DRGLI1       ASCFSTPRSM LKLSKKRALS ISP.LSDASV DLQTVIRTSP NSLVAF...V
AMPHIGLI     ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
DMCI         RPPGGSIRA. .SISRKRALS SSP.YSD.SF DINSMIRFSP NSLAT...IM

             251                                                300
CIONAGLI     NSRSSSVASG GSYGHLAAGG IS.SI.FSS. .....KPTY. ..KVVLYTFI
DRGLI1       NSRCGP.NNP SSYGHLSVGT MSPSLGFSS. .....SINYS RPQGNIYSHP
AMPHIGLI     ~~~~~~~AST GSYGHLSASA MRTESGAES. .....KPGDP VLRKHAVQRA
DMCI         NGSRGSSAAS GSYGHISATA LNPMSHVHST RLQQIQAHLL RASAGLLNPM
```

A) BLOSUM62



C) BLOSUM85



C) BLOSUM100

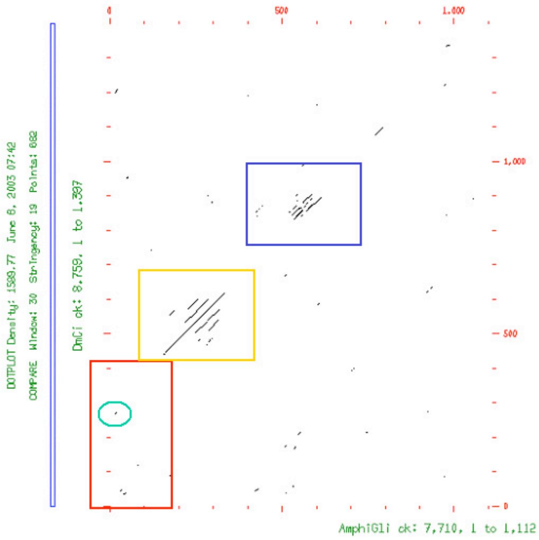


Figure 2: Graphical alignments of AmphiGli and DmCi with various scoring matrices.

The Graph function on GCG SeqWeb was used to compare the sequences of AmphiGli and DmCi. Red boxes highlight N-terminal sequences containing the SuFu motif. Orange boxes highlight the zinc fingers. Blue boxes highlight the PKA phosphorylation motif region in the C-terminal half of each protein. **A**) Alignment with the BLOSUM62 scoring matrix demonstrates a high degree of global amino acid similarity in the Ci/Gli family proteins and provides a possible explanation for the failure of multiple alignment algorithms to correctly align these sequences. **B and C)** Increasing the stringency of the scoring matrix (to BLOSUM85 and BLOSUM100 respectively)reduces the number of possible diagonals from which an alignment algorithm must choose to construct an alignment. Increasing the window size from the default of 30 to the maximum of 50 has a similar effect to using a BLOSUM100 matrix. In **(C)**, the green circle highlights a diagonal containing the SuFu motif. This alignment was confirmed using the Gap tool (Needleman-Wunsch algorithm) in GCG SeqWeb **(D)**.

D) Needleman-Wunsch alignment of AmphiGli and DmCi

```
Needleman-Wunsch (BLOSUM62)GOP 8, GEP 2
        1 ..................................ASTGSYGHLSASAMRT 16
                                            || |    :.   .
      251 aasgsyghisatalnpmshvhstrlqqiqahllrasagllnpmtpqqvaa 300

Needleman-Wunsch (BLOSUM100) GOP 8, GEP 2
        1 ASTGSYGHLSASAMRTESGAESKPGDPVLRKHAVQ....RADAHVPVPTS 46
          |..|||||.||.|:   |   |       | . :|    || |    | .
      251 aasgsyghisatalnpmshvhst......rlqqiqahllrasagllnpmt 294
```

Figure 3: ClustalW alignment of N-terminus with Identity matrix

```
AmphiGli        ------ASTGSYGHLSASAMRTESGAESKPGDPVLRKHAVQRADAHVPVPTS-------- 46
DmCi            MKWTPTRYLHIFLLPSRRAAAVAAAATVLPGSPCINQHHPTDVSSSVTVPSIIPTGGTSD 60

AmphiGli        -------------------------------------------------PAMQQFHNRLM 57
DmCi            SIKTSIQPQICNENTLLGNAGHQHNHQPQHVHNINVTGQPHDFHPAYRIPGYMEQLYSLQ 120

AmphiGli        RQKSPFHFGMPHASPFAAPLPAGM---------------------AMLAAQGAM------ 90
DmCi            RTNSASSFHDPYVNCASAFHLAGLGLGSADFLGSRGLSSLGELHNAAVAAAAAGSLASTD 180

AmphiGli        ---------------PPSSSAATHTETKAGEPSS-------------------------- 109
DmCi            FHFSVDGNRRLGSPRPPGGSIRASISRKRALSSSPYSDSFDINSMIRFSPNSLATIMNGS 240

AmphiGli        ---------------------SIVSST--------------------------------- 115
DmCi            RGSSAASGSYGHISATALNPMSHVHSTRLQQIQAHLLRASAGLLNPMTPQQVAASGFSIG 300
```

Figure 4A: ClustalW alignment of C-terminus, Negative Matrix Values OFF

```
AmphiGli        HKPTGQTCDAQQSVYGSSPHHDSGVEMNANS-GSLPDLSTLDDQVISDSSISSTVPTSRA 441
CionaGli        ----------------TSQNNDSGVDVNVGG--------NEGDSDGDIVVDENPQPDSTS 390
DmCi            DISSSNHHLVNGVRASDSLLTYSPDDLAENL-NLDDGWNCDDDVDVADLPIVLRAMVNIG 718
DrGli1          ------SCSSERSPLGSANNNDSGVEMNLNAAGSLEDLTTQEDSGNAGVSESSATISS-- 567

AmphiGli        SGVMVAARPGLVPRAPRIGNKPSNQRRRMRLSSGTPGPTSPPRSDSVQLPPIEKTGSRGP 501
CionaGli        GGVGVQSRH---RGTVRASMVPRLVNKKMQNLSLGGLSPNVE------------------ 429
DmCi            NGNASASTIGGSVLARQRFRGRLQTKGINSSTIMLCNIPESNRTFGISELNQRITELKME 778
DrGli1          GGMCMSVQA--LKRLENLKIDKLKQIRRPTPPGRNAGNKLPALSATGEMMSMCAPSPLLS 625

AmphiGli        SAQGSHSSVEAANRRTNELRASDLSQTSRTSSLGSLGSRKDSASTVSSYYSSRRSSEASP 561
CionaGli        ------SYV-DIGGYDDQRKLGNFTEVSSTTAFPAKQKSTTYPRKLPLTPHRQVALLNQD 482
DmCi            PGTDAEIKIPKLPNTTIGGYTEDPLQNQTSFRNTVSNKQGTVSGSIQGQFRRDSQNSTAS 838
DrGli1          NRRVMELSAPDMGGVTGMSCPPNDRRGSGTSSLSSAYTVSRRSSMVSPYLSSRRSSDVSH 685

AmphiGli        FPESIFSSRRSSQASPFPGINRRTSNGSLYSPNDSYDPISLGSSRKSSDASSLSMNVNEL 621
CionaGli        RRDSGTVSDGSRKSSMASQNSRRSSQNTGFNVAGSYDPISLDSSRRSSANCGSG------ 536
DmCi            TYYGSMQSRRSSQSSQVSSIPTMRPNPSCNSTASFYDPISPGCSRRSSQMSNGAN-C--- 894
DrGli1          CQSVMGGEVPGDPLSPQNSQRAGLCQNSGGLPGLPSLTPAQQYSLKAKYAAATGGPPPTP 745

AmphiGli        GINIEQQQMLRARFIQATGRPPTAVCGNDSRPESRRGDRKEKENVEEPNPRRQSDLGHYN 681
CionaGli        ---------------------SSTINAFHLHRLRSRFNEDAGLPPPTPLDREGYTKSQLS 575
DmCi            -----NSFTSTSGLPVLNKESNKSLNACINKPNIGVQGVGIYNSSLPPPPSSHLIATNLK 949
DrGli1          LPNMDQAGTPARHVGFLRECQGQPLPPFLQQGGTRRHSANAEYGTGVIYPHQAPGNNTRR 805

AmphiGli        RLKGTPLPKEVKDGPHRRSSAPQKNDVVTNLPDVPRDHSFNKHTPLPPVTPQPPPQIKKA 741
CionaGli        RWFKDEQPTVDPAGYQFNPQARPSLPQMGPPKTPEVRRRSEGAQSRPSRTPLPQHLGGNA 635
DmCi            RLQRKDSEYHNFTSGRFSVPSYMHSLHIKNNKPVGENEFDKAIASNARRQTDPVPNINLD 1009
DrGli1          ASDPVRSAADPQGLPKVQRFNSLSNVSLMSRRNALQQCGSDAALSRHMYSPRPPSITENV 865

AmphiGli        FSPSKVKQAFSPKSASTSMQGVAEEFPMDLIENEPDVIIPDEMVQFLNSQTGDDPREMVP 801
CionaGli        FRRASD------------------------------------------------------ 641
DmCi            PLTNISRFSTTPHSFDINVGKTNNIASSINKDNLRKDLFTVSIKADMAMTSDQHPNERIN 1069
DrGli1          MMEAMGMDGNTEGRQQGNMIPGGDRSYMGYQHNPHQASQLSPGQESLGCIDQVYQSQMQG 925
```

Figure 4B: ClustalW alignment of C-terminus, Negative Matrix Values ON

```
AmphiGli        NGVHSSTTNPAA---SQGSPGQKPTEGHKPTGQTCDAQQSVYGSSPHHDSGVEMNANS-G 414
CionaGli        -----------------------------------TSQ--------NNDSGVDVNVGGNE 371
DmCi            Q-EHNIDSSPCSEDSHLGKMLGTSSPSIKSESDISSSNHHLVNGVRASDSLLTYSPDD-- 685
DrGli1          NREDCKLLAPDNTLKSQPSPGGQSS---------CSSERSPLGSANNNDSGVEMNLNAAG 542

AmphiGli        QTSRTSSLGSLGSRKDSASTVSSYYSSRRSSEASPFPESIFSSRRSSQASPFPGINRRTS 586
CionaGli        -----S---------------TTAFPAKQKSTTYPRKLPLTPHRQVALLNQDRRDSGTVS 490
DmCi            NQTSFRNTVSNKQGTVSGSIQGQFRRDSQNSTASTYYGSM-QSRRSSQSSQVSSIPTMRP 863
DrGli1          GVTGMSCPPNDRRGSGTSSLSSAYTVSRRSSMVSPY---L-SSRRSSDVSHCQSVMGGEV 694

AmphiGli        NGSL-----------------YSPNDSYDPISLGSSRKSSDASSLSMNVNELGINIEQQQ 629
CionaGli        DGSRKSSMASQNSRRSSQNTGFNVAGSYDPISLDSSRRSSANCGSGS-S---TINAFHLH 546
DmCi            NPSCNST---------------ASF--YDPISPGCSRRSSQMSNGANCNS---------- 896
DrGli1          PG--------------------------DPLSPQNSQRAGLCQNSGGLPGLPSLTPAQQY 728

AmphiGli        MLRARFIQATGRP-------------------------------PTAVCGNDSRPESR-- 656
CionaGli        RLRSRFNEDAGLP-------------------------------PPTPLDREGYTKSQLS 575
DmCi            -----FTSTSGLPVLNKESNKSLNACINKPNIGVQGVGIYNSSLPPPPSSHLIATNLKRL 951
DrGli1          SLKAKYAAATGGP-------------------------------PPTPLPNMDQAGT--- 754

AmphiGli        --------------------RGDRKEKENVEEPNPRRQSDLGHYNRLKGTPLPKEVKDGP 696
CionaGli        RWFKDEQPTVDPAGYQFNPQARPSLPQMGPPKTPEVRRRSEGAQSRPSRTPLPQHLGGNA 635
DmCi            QRKDSEYHNFTSGRFSVPSYMHSLHIKNNKPVGENEFDKAIASNARRQTDPVP------- 1004
DrGli1          ----------------------P--ARH------------VGFLRECQGQPLPPFLQQGG 778

AmphiGli        HRRSSAPQKNDVVTNLPDVPRDHSFNKHTPLPPVTPQPPPQIKKAFSPSKVKQAFSPKSA 756
CionaGli        FRRASD------------------------------------------------------ 641
DmCi            -NINLDPLTN---------ISRFSTTPHSFDINVGKTNNIASSINKDNLRKDLFTVSIKA 1054
DrGli1          TRRHSANAEYGTGVIYPHQAPGNNTRRASDPVRSAADPQGLPKVQRFNSLSNVSLMSRRN 838
```

Figure 4C: T-Coffee Alignment of C-terminus

```
AmphiGli       SVYGSSPHHD SGVEMNANS. GSLPDLSTLD ..DQVISDSS ISSTVPTSRA
CionaGli       ....TSQNND SGVDVNVGGN EGD....... SDGDIVVDEN PQ...PDSTS
DmCi           SIKSESDISS SNHHLVNGVR ASD....... SLLTYSPDDL AE...NLNLD
DrGli1         SPLGSANNND SGVEMNLNAA GSLEDLTTQE DSGNAGVSES SA...TIS.S

AmphiGli       SGVMVA.... .......... ...ARPGLVP RAPRIGNKPS NQRRRMRLSS
CionaGli       GGVGVQSRHR GTVRASMVPR LVNKKMQNLS LGGLSPNV.. ..........
DmCi           DGWNCDDDVD VADLPIVLRA MVNIGNGNAS ASTIGGSVLA RQRFRGRLQT
DrGli1         GGMCMSVQAL KRLENLKIDK LKQIRRPTPP GRNAGNKL.. ..........

AmphiGli       .GTPGPT... .SPPRSD.SV QLPPIEKTGS RGPS...... ..........
CionaGli       ....ESYVDI GGYDDQRKLG NFTEVSSTTA FPAKQKSTTY PRKLPL....
DmCi           KGINSSTIML CNIPESNRTF GISELNQRIT ELKMEPGTDA EIKIPKLPNT
DrGli1         ....PALSAT GE........ .......... .......... ..........

AmphiGli       .AQGSHSSVE AANRRTNELR ASDL...... ...SQTSRTS SLGSLGSRKD
CionaGli       .......... .......... ....TPHRQV ALLNQDRRDS GTVSDGSRKS
DmCi           TIGGYTEDPL QNQTSFRNTV SNK.QGTVSG SIQGQFRRDS QNSTASTYYG
DrGli1         MMSMCAPSPL LSNRRVMELS APDMGGVTGM SCPPNDRRGS GTSSLSS..A

AmphiGli       SASTVSSYYS SRRSSEASPF PESIFSSRRS SQASPFPGIN RRTSNGSLYS
CionaGli       SM........ .......... .....ASQNS RRSSQ..... ...NTG..FN
DmCi           SM........ .......... .....QSRRS SQSSQVSSIP TMRPNPSCNS
DrGli1         YT........ .......... .....VSRRS SMVSPYLSSR RSSDVSHCQS

AmphiGli       PNDSY...DP ISLGSSRKSS DAS......S LSMNVNELGI NIEQQQMLRA
CionaGli       VAGSY...DP ISLDSSRRSS .....ANCGS GSSTINAFHL HRLRSRFNE.
DmCi           TASFY...DP ISPGCSRRSS QMSNGANCNS FTSTSGLPVL NKESNKSLNA
DrGli1         VMGGEVPGDP LSPQNSQRAG L......CQN SGGLPGLPSL TPAQQYSLKA

AmphiGli       RFIQATGRPP TAVCGNDSRP ES........ .......... ...RRGDRKE
CionaGli       .......... .....DAGLP PP.......T PLDR...... ..........
DmCi           CINKPNIGVQ GVGIYNSSLP PPPSSHLIAT NLKR...... .LQRKDSEYH
DrGli1         KYAAATGGPP PTPLPNMDQA GTPARHVGFL RECQGQPLPP FLQQGGTRRH

AmphiGli       KENVE..... .......... .......... .......... ......EPNP
CionaGli       .......... .......... .......... .......... ..........
DmCi           NFTSGRFSVP SYMHSLHIKN NKPVGENEFD KAIASNARRQ TDPVPNI...
DrGli1         SANAEYGTGV IYPH...... .......... QAPGNNTRRA SDPVRSAADP

AmphiGli       RRQSDLGHYN RLKGTPLPKE VKDGPHRRSS APQKNDVVTN LPDVPRDHSF
CionaGli       .......... .......... .......... .......... ..........
DmCi           .......NLD PLTNISRFST T......... .......... ..........
DrGli1         QGLPKVQRFN SLSNVSLMSR RNALQQCGSD AALSRHMYSP RPPSITENVM

AmphiGli       NKHTPLPPVT P......... .......QPP PQIKKAFSPS KVKQAFSPKS
CionaGli       .......... .......... .......... .......... ..........
DmCi           .......... .........P HSFDINVGKT NNIASSINKD NLRKDLFTVS
DrGli1         MEAMGMDGNT EGRQQGNMIP GGDRSYMGYQ HNPHQASQLS PGQESLGCID

AmphiGli       ASTSMQGVAE EFPMDLI..E NEPDVIIPDE MVQFLN.... ....SQTGDD
CionaGli       .......... .......... .......... .......... ..........
DmCi           IKADMAMTSD QHPNERINLD EVEELILPDE MLQYLNLVKD DTNHLEKEHQ
DrGli1         QVYQSQMQGQ YQREESCSTG VMGQADIANN LLQQAEYGMS TCQLSPSGPH

AmphiGli       PREMVPNFEQ VGTTPTFVED IPPMQVNPIQ GDGFSNMGSP QQAFSPNRQP
CionaGli       .....EGYTK SQLSRWFKDE QPTVD...PA GYQFNPQARP S..LPQMGPP
DmCi           AVPVGSNVSE TIASNHYREQ SNIYY...TN KQILTPPSNV D..IQPNTTK
DrGli1         YPSQGDGSGP WGQTNQLHSP GMQYQGAGMQ GQHYTQQGIY DPTSNPNLQR

AmphiGli       MP........ .......... .......... ......PIQQ QQAFNQSQQV
CionaGli       KTPEVRRRSE GAQSRPSR.. .......... .......... ..........
DmCi           FTVQDKFAMT AVGGSFSQRE LSTL...... .......... .AVPNEHGHA
DrGli1         VTVKPEQFHP SMGGSSSCQN TKALHQNRHN ANMQTYPLQG QGIMNRSSSA
```

Figure 4D: DiAlign Alignment of C-terminus

```
AmphiGli  359   ngvhsstt-- NPAASQGSPG qkpteghkpt gqTCDAQQSV YGSSPHHDSG
CionaGli  355   ---------- ---------- ---------- ---------- --TSQNNDSG
DmCi      613   ---------- ---------- ---------- ----NDANSR LQQNNSRHNL
DrGli1    493   redckllapd NTLKSQPSPG gqs------- --SCSSERSP LGSANNNDSG

AmphiGli  407   VEMNAN-SGS LPDLSTLDDQ VISDSSISST VPTSRASGVM VAARPGLVPR
CionaGli  363   VDVNVggneg dsdgdivvde npq------- -PDSTSGGVG VQSRHRGTVR
DmCi      629   QEHNIDSSPC SEDshlgkm- ---------- ---------- ----------
DrGli1    534   VEMNLNAAGS LEDLTTQEDS GNAGVSESSA TISSGGMCMS VQAlkrlenl

AmphiGli  526   ---SQTSRTS SLGSLGSRKD SASTVSSYYS SRRSSEASpf pesifssRRS
CionaGli  480   ---NQDRRDS GTVSDGSRKS SMASQNSRRS SQNTgfnv-- ----------
DmCi      813   VSNKQGTVSG SIQGQFRRDS QNSTASTYYG SMQS------ -------RRS
DrGli1    647   PNDRRGSGTS SLSSAYTVSR RSSMVSPYLS SRRSSDVShc qsvmggevpg

AmphiGli  573   SQASPFPGIN RRTSNGSLYS PNDSYDPISL GSSRKSSdas slsmnvnelg
CionaGli  515   ---------- ---------- -AGSYDPISL DSSRRSS--- --ANCGSGSS
DmCi      850   SQSSQVSSIP TMRPNPSCNS TASFYDPISP GCSRRSSqms ngANCNSFTS
DrGli1    697   ---------- ---------- -----DPLSP QNSQRA---- ----------

AmphiGli  623   inie------ ---------- ---------- ---------- ----------
CionaGli  539   Tinafhlhrl rsrfnedag- ---------- -----LPPPt pldregytks
DmCi      900   Tsglpvlnke snkslnacin kpnigvqgvg iynssLPPP- ----------
DrGli1    708   ---------- ---------- ---------- ---------- ----------

AmphiGli  627   ---------- ---------- ---------- ----QQQMLR ARFIQATGRP
CionaGli  573   qlsrwfkdeq ptvdpaGYQF NPQARPSLPQ MGPPKTPEVR RRSEGAQSRP
DmCi      939   ---------- ---------- ---------- ---------- ----------
DrGli1    708   ---------- ------GLCQ NSGGLPGLPS LTPAQQYSLK AKYAAATGGP

AmphiGli  643   Ptavcgndsr pesrrgdrke kenveepnpr rqsdLGHYNR LKGTPLPKEV
CionaGli  623   SRTPLPqhlg ---------- ---------- ---------- ----------
DmCi      939   ---------- ---------- ---------- ---------- ----------
DrGli1    742   PPTPLPnmdq agtparh--- ---------- ----VGFLRE CQGQPLPPFL

AmphiGli  693   KDGPHRRSSA PQKNDVVTNL PDVPRDHSfn khtplppvtp qppp------
CionaGli  633   ---------- ---------- ---------- ---------- ----------
DmCi      939   ---------- PSSHLIATNL KRLQRKDSey hnftsgrfsv psymhslhik
DrGli1    775   QQGGTRRHSA naeygtgviy ph-------- ---------- ----------

AmphiGli  737   ---------- ---------- ---------- ---------- ----------
CionaGli  633   ---------- ----GNAFRR ASD------- ---------- ----------
DmCi      979   nnkpvgenef dKAIASNARR QTDPVpninl dpltnisrfs ttphsfdinv
DrGli1    797   ---------- -QAPGNNTRR ASDPVrsaad pqglpkvqrf nslsnvslms
```

Figure 4E: Pileup Alignment of C-terminus

```
CionaGli SLRKHVKTVH GPAAHVTKRM KM...TSQNN DSGVDVNVGG N.E.......
  DrGli1 slrkhvktvh gpeahitkkh rg...dtgpr ppglttagqs s.elliekee
AmphiGli SLRKHVKTVH GPEAHQTKKH KTLGPTPRPR DPPSEKRDQD SVSSPPDSNG
    DmCi slrkhvktvh gaefyankkh kgl.....pl ndansrlqqn nsrhnlqehn

         651                                               700
CionaGli .GDSDGDIVV DEN.....PQ P......... .......... DSTSGGVGVQ
  DrGli1 rnredcklla pdntlksqps pggq...ssc ssersplgsa nnndsgvemn
AmphiGli VHSSTTNPAA SQGSPGQKPT EGHKPTGQTC DAQQSVYGSS PHHDSGVEMN
    DmCi idsspcseds hlgkmlgtss psiksesdis ssnhhlvngv rasdslltys

         801                                               850
CionaGli YPRKLPLTPH RQVALLN... .......... .QDRRDSGTV S.....DGSR
  DrGli1 mcapspllsn rrvmelsapd mggvtgmscp pndrrgsgts slssaytvsr
AmphiGli GSHSSVEAAN RRTNELRASD LSQTSRTSSL GSLGSRKDSA STVSSYYSSR
    DmCi aeikipklpn ttiggytedp lqnqtsfrnt vsnkqgtvsg siqgqfrrds

         851                                               900
CionaGli KSSMASQ... ...NSRRSS. ....QNT.GF NVAG...... .SYDPISLDS
  DrGli1 rssmvspy.. ..lssrrssd vshcqsvmgg evpg...... ...dplspqn
AmphiGli RSSEASPFPE SIFSSRRSSQ ASPFPGINRR TSNGSLYSPN DSYDPISLGS
    DmCi qnstastyyg s.mqsrrssq ssqvssiptm rpnpscnsta sfydpispgc

         901                                               950
CionaGli SRRSS..ANC G......S.. GSSTINAFHL HRLRSRFN.. ....EDAG..
  DrGli1 sqraglcqns g......glp glpsltpaqq yslkakya.. ....aatg..
AmphiGli SRKSSDASSL S......MNV NELGINIEQQ QMLRARFI.. ....QATG..
    DmCi srrssqmsng ancnsftsts glpvlnkesn kslnacinkp nigvqgvgiy

         951                                              1000
CionaGli ..LPPPTPLD REGYTKSQL. SRWFKD.... EQPT.VDP.. AGYQFNPQAR
  DrGli1 ..gppptplp n......... ....md.... qagtparh.. vgflrecqgq
AmphiGli ..RPPTAVCG NDSRPESRRG DRKEKE..NV EEPNPRRQSD LGHYNRLKGT
    DmCi nsslppppss hliatnlkrl qrkdseyhnf tsgrfsvpsy mhslhiknnk

         1001                                             1050
CionaGli PSLPQMGPPK TPEVRRRSEG AQSRPSRTPL PQHLGGNAFR RASD~~~~~~
  DrGli1 plppflqqgg t...rrhsan ae.ygtgviy phqapgnntr rasdpvrsaa
AmphiGli PLPKEVKDG. ...PHRRSSA PQKNDVVTNL PDVPRDHSFN KHTPLPPVTP
    DmCi pvgenefdka iasnarrqtd pvpninldpl tnisrfsttp hsfd.invgk
```

Figure 5: Pairwise BLAST alignments of Ci/Gli sequences.

5A.

```
  Query: AmphiGli
  Subject: CionaGli

Query: 359 NGVHSSTTNPAASQGSPGQKPTEGHKPTGQTCDAQQSVYGSSPHHDSGVEMN--ANSGSL 416
                                                   +S ++DSGV++N   N G
Sbjct: 355 ----------------------------------------TSQNNDSGVDVNVGGNEGDS 374

Query: 537 LGSRKDSASTVSSYYSSRRSSEASPFPESIFSSRRSSQASPFPGINRRTSNGSLYSPNDS 596
              R+DS  TVS    SR+SS AS       +SRRSSQ + F             +   S
Sbjct: 482 --DRRDS-GTVSD--GSRKSSMASQ------NSRRSSQNTGF-------------NVAGS 517

Query: 597 YDPISLGSSRKSS-DASSLSMNVNELGINIEQQQMLRARFIQATGRPPTAVCGNDSRPES 655
           YDPISL SSR+SS +  S S  +N   ++      LR+RF +  G PP      +   +S
Sbjct: 518 YDPISLDSSRRSSANCGSGSSTINAFHLH-----RLRSRFNEDAGLPPPTPLDREGYTKS 572

Query: 656 R--RGDRKEKENVE---------------------EPNPRRQSDLGHYNRLKGTPLPKEV 692
           +  R  + E+  V+                      P  RR+S+ G  +R   TPLP+ +
Sbjct: 573 QLSRWFKDEQPTVDPAGYQFNPQARPSLPQMGPPKTPEVRRRSE-GAQSRPSRTPLPQHL 631

Query: 693 KDGPHRRSS 701
                RR+S
Sbjct: 632 GGNAFRRAS 640
```

5B.

```
  Query: AmphiGli
  Subject: DmCi

Query: 371  SQGSPGQKPTEGHKPTGQTCDAQQSVYGSSPHHDSGVEMNANSG----------SLP--- 417
               S     +  H   G    A  S+   SP  D    +N + G           LP
Sbjct: 656  KSES-DISSSNHHLVNG--VRASDSLLTYSP-DDLAENLNLDDGWNCDDDVDVADLPIVL 711

Query: 523  SDLSQTSRTSSLGSLGSRKDSASTVSSYYSSRRSSEASPFPESIFSSRRSSQASPFPGIN 582
                Q + + S+     R    ST S+YY S               SRRSSQ+S    I
Sbjct: 817  ----QGTVSGSIQGQFRRDSQNSTASTYYGS-------------MQSRRSSQSSQVSSIP 859

Query: 583  RRTSNGSLYSPNDSYDPISLGSSRKSSDASSLSMNVNELG-------INIEQQQMLRARF 635
                N S  S    YDPIS G SR+SS  S+   N N          +N E  + L A
Sbjct: 860  TMRPNPSCNSTASFYDPISPGCSRRSSQMSN-GANCNSFTSTSGLPVLNKESNKSLNA-- 916

Query: 636  IQATGRPPTAVCG----NDSRP----------ESRRGDRKEKE-------NVEEPNPRRQ 674
                 +P   V G    N S P            +R  RK+ E           P+
Sbjct: 917  --CINKPNIGVQGVGIYNSSLPPPPSSHLIATNLKRLQRKDSEYHNFTSGRFSVPSYMHS 974

Query: 675  SDLGHYNRLKGTPLPKEVKDGPHRRSSAPQKNDVVTNLPDVPR----DHSFNKHTPLPPV 730
              + +   +      K +     RR + P  N  +  L ++ R     HSF+       +
Sbjct: 975  LHIKNNKPVGENEFDKAIASNA-RRQTDPVPNINLDPLTNISRFSTTPHSFD-------I 1026
```

5C.

```
  Query: AmphiGli
  Subject: DrGli1

Query: 355  PPDSNGVHS---STTNPAASQGSPGQKPTEGHKPTGQTCDAQQSVYGSSPHHDSGVEMNA 411
              + N       +  N   SQ SPG +          +C +++S  GS+ ++DSGVEMN
Sbjct: 488  KEERNREDCKLLAPDNTLKSQPSPGGQ---------SSCSSERSPLGSANNNDSGVEMNL 538

Query: 531  TSSLGSLGSRKDSASTVSSYYSSRRSSEASPFPESIFSSRRSSQASPFPGINRRTSNGSL 590
             S   +      ++S  S+Y  SRRSS  SP+     SSRRSS  S    +      G
Sbjct: 643  MSCPPNDRRGSGTSSLSSAYTVSRRSSMVSPY----LSSRRSSDVSHCQSVMGGEVPGDP 698

Query: 591  YSPNDSYDPISLGSSRKSSDASSLSMNVNELGINIEQQQMLRARFIQATGRPPTAVCGND 650
             SP +S      G  + S     L        +   QQ  L+A++  ATG PP     N
Sbjct: 699  LSPQNSQ---RAGLCQNSGGLPGLP------SLTPAQQYSLKAKYAAATGGPPPTPLPNM 749

Query: 651  SRPESRRGDRKEKENVEEPNPRRQSDLGHYNRLKGTPLPKEVKDGPHRRSSAPQKNDVVT 710
                            +   P R   +G     +G PLP  ++ G  RR SA  +
Sbjct: 750  D---------------QAGTPARH--VGFLRECQGQPLPPFLQQGGTRRHSANAEYGTGV 792

Query: 711  NLP-DVPRDHSFNKHTPL-PPVTPQPPPQIKKAFSPSKVKQAFSPKSASTSMQGVAEEFP 768
              P   P +++     P+     PQ  P++++  S S V    S  S   ++Q    +
Sbjct: 793  IYPHQAPGNNTRRASDPVRSAADPQGLPKVQRFNSLSNV----SLMSRRNALQQCGSDAA 848
```

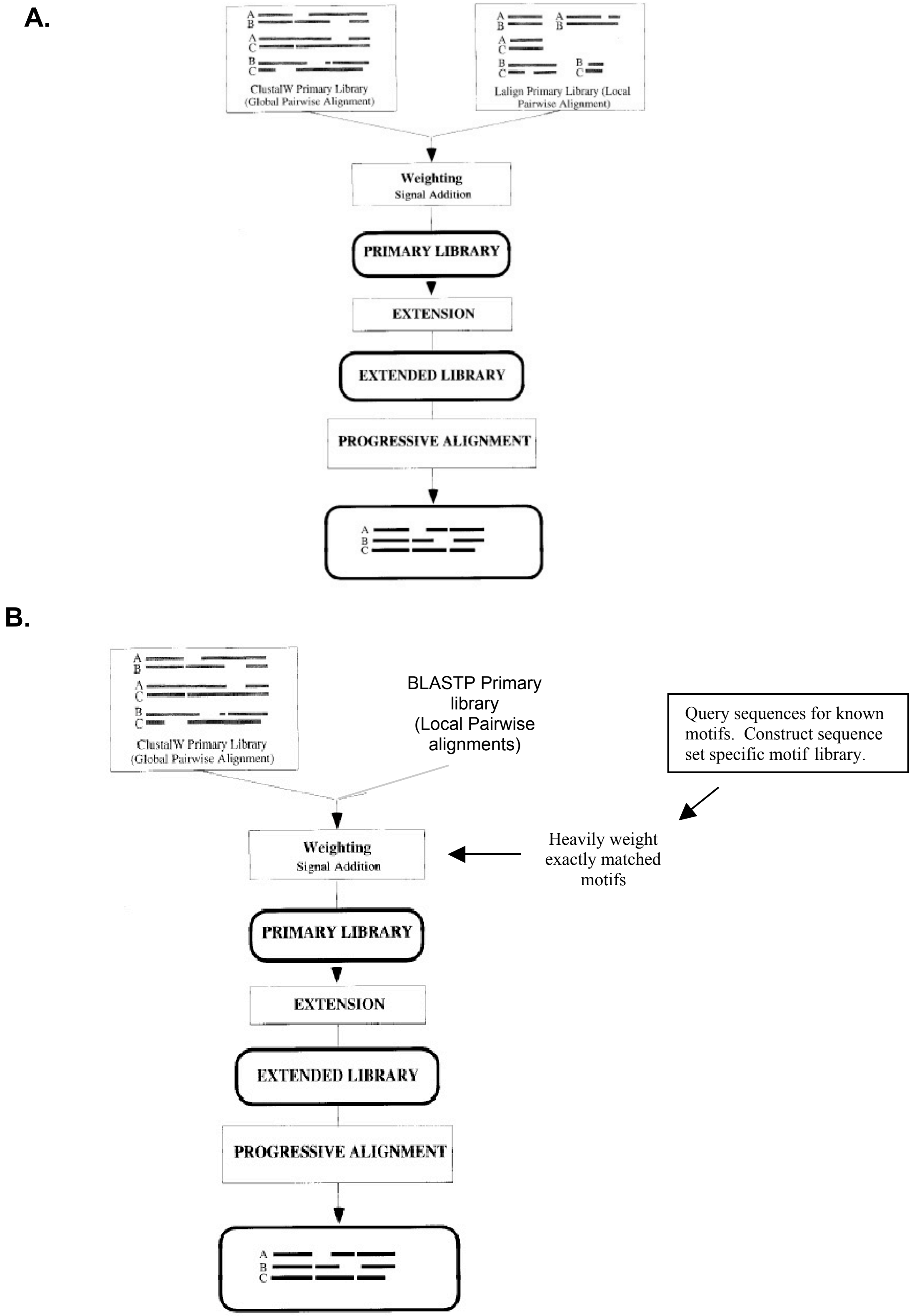
5D.

```
  Query:CionaGli
  Subject: DrGli1

Query: 354 -----MTSQNNDSGVDVNVGGNEGDSDGDIVVDENPQPDSTSGGVGVQSRHRGTVRASMV 408
                 ++ NNDSGV++N+  N   S  D+   E+      SG  GV S    T+ +  +
Sbjct: 520 ERSPLGSANNNDSGVEMNL--NAAGSLEDLTTQED------SGNAGV-SESSATISSGGM 570

Query: 467 KLPLTPHRQVALLN--------------QDRRDSGTVS-----DGSRKSSMASQ--NSRR 505
             PL  +R+V  L+               DRR SGT S       SR+SSM S   +SRR
Sbjct: 622 --PLLSNRRVMELSAPDMGGVTGMSCPPNDRRGSGTSSLSSAYTVSRRSSMVSPYLSSRR 679

Query: 506 SSQNT------GFNVAGSYDPISLDSSRRSS--ANCGS--GSSTINAFHLHRLRSRFNED 555
           SS  +      G  V G  DP+S  +S+R+    N G   G  ++     + L++++
Sbjct: 680 SSDVSHCQSVMGGEVPG--DPLSPQNSQRAGLCQNSGGLPGLPSLTPAQQYSLKAKYAAA 737

Query: 556 AGLPPPTPLDREGYTKSQLSRWFKDEQPTVDPA-------GYQFNPQARPSLPQMGPPKT 608
            G PPPTPL                  P +D A       G+    Q +P  P +    T
Sbjct: 738 TGGPPPTPL------------------PNMDQAGTPARHVGFLRECQGQPLPPFLQQGGT 779

Query: 609 PEVRRRSEGAQSRPSRTPLPQHLGGNAFRRASD 641
              RR S  A+   +    P    GN  RRASD
Sbjct: 780 ---RRHSANAE-YGTGVIYPHQAPGNNTRRASD 808
```

5E.

```
  Query:DmCi
  Subject: DrGli1

Query: 676 DSLLTYSPDDLAENLNLDDGWNCDDDVDVADLPIVLRAMVNIGNGNASASTIGGSVLARQ 735
                 + +D    +NL+   + +D     D         N G   +SA+   G +
Sbjct: 528 ------NNNDSGVEMNLNAAGSLEDLTTQED-------SGNAGVSESSATISSGGMCMSV 574

Query: 791 PNT--TIGGYTEDPLQNQTSFRNTVSNKQGTVSG-SIQGQFRRDSQNSTASTYYGSMQSR 847
             T   +      PL +        +   G V+G S     RR S  S+ S+ Y    SR
Sbjct: 609 SATGEMMSMCAPSPLLSNRRVMELSAPDMGGVTGMSCPPNDRRGSGTSSLSSAY--TVSR 666

Query: 848 RSSQSSQVSSIPTMRPNPSCNSTASFY---DPISPGCSRRSSQMSNGANCNSFTSTSGLP 904
           RSS  S   S         C S        DP+SP  S+R+        C +     GLP
Sbjct: 667 RSSMVSPYLSSRRSSDVSHCQSVMGGEVPGDPLSPQNSQRAGL------CQNSGGLPGLP 720

Query: 905 VLNKESNKSLNACINKPNIGVQGVGIYNSSLPPPPSSHL 943
            L      SL A       G     + N      P+ H+
Sbjct: 721 SLTPAQQYSLKAKYAAATGGPPPTPLPNMDQAGTPARHV 759
```

Figure 6: T-Coffee algorithm (A) and proposed modifications (B).


A.
A
B
A
C
B
C
ClustalW Primary Library
(Global Pairwise Alignment)
A
B
A
B
A
C
B
C
B
C
Lalign Primary Library (Local Pairwise Alignment)
Weighting
Signal Addition
PRIMARY LIBRARY
EXTENSION
EXTENDED LIBRARY
PROGRESSIVE ALIGNMENT
A
B
C
B.
A
B
A
C
B
C
ClustalW Primary Library
(Global Pairwise Alignment)
BLASTP Primary library
(Local Pairwise alignments)
Query sequences for known motifs. Construct sequence set specific motif library.
Heavily weight exactly matched motifs
Weighting
Signal Addition
PRIMARY LIBRARY
EXTENSION
EXTENDED LIBRARY
PROGRESSIVE ALIGNMENT
A
B
C