A Bioinformatics Investigation of Cysteine-Rich PAK1 Inhibitor (CRIPak)

Introduction

The cysteine-rich PAK1 inhibitor (CRIPak) is a known inhibitor of p21-activated protein kinase 1 (Pak1), composed of multiple zinc finger and trypsin-inhibitor like cysteine-rich domains. CRIPak is an intronless gene, localized on chromosome 4p16.3, with an open reading frame of 446 amino acids. Its demonstrated interactivity with the Pak1 protein involves amino acids 367-447 in the C-terminus (Talukder et. al). Little is known of CRIPak's gene family, or of its phylogenetic history in regards to other related genes.

The evolutionary research done on the CRIPak inhibitor is minimal, if that. Of that research that has been conducted, most involves the evolution of the CRIPak gene in tumorigenic tissue. Copy number loss and decreased mRNA expression in CRIPak were found to have significant correlation with aggressive tumors in both lung (Qian et. al) and breast (Gonzalez-Angulo et. al) cancer tissues. Investigation into follicular cancer also found a predominant trend of recurrent genetic mutations in the CRIPak gene (Asmann et. al). While certainly interesting, not much of the established research on CRIPak can be reconciled in terms of its relationship to other homologous genes, or *in regard to* its phylogenetic history. Some research has been done on the evolution of the PAK gene family, a primary target of the CRIPak protein, specifically in its evolution in early metazoa. The CRIB-AID motif, an area largely active in CRIPak interaction, has shown a dramatic shift in identity and alignment in comparing animal and non-animal sequences, specifically in regulatory domains (Watari et. al). These results might suggest either a similarly dramatic shift in CRIPak evolution, or at the very least some genetic change during a similar time period. While we anticipate this event going into our analysis, this is essentially all of the preliminary information available at present.

Ultimately, in this paper, we hope to further the scientific understanding of CRIPak evolution, specifically in considering its evolutionary patterns between cnidaria and bilateria, either by independent convergence, or divergence from a common ancestor. By conducting various bioinformatics analyses, we hope to compile evidence outlining a proposed genetic history of CRIPak and related genes, defining various speciation and gene duplication events that will aid in a telling of its evolutionary story. This will hopefully provide some insight into the functional evolution of the proteins encoded by this gene/gene family, and may even shed some light on the role of CRIPak in cancer and oncogenetics.

Methods

Sequence analysis begins with the exploration of putative homologs, from which we eventually hope to further extrapolate a phylogenetic representation of gene family evolution. We first perform a BLAST search (Altschul et. al) of the NCBI CRIPAK query sequence against a database of given proteomes, filtering search results to only include those putative homologs with an e-value of less than 10^{-10}. From here, the filtered BLAST file can be manipulated into a fasta file using SEQKIT (Shen et. al) and produce an alignment of resultant homologs, using muscle (Edgar) for the alignment and t_coffee (Notredame et. al) for alignment statistics. Alv (Arvestad) can then be used to visualize the alignment on the command line. Highly gapped regions can also be removed using t_coffee, allowing for a better visual representation of sequence alignment.

From these aligned sequences, we can begin to hypothesize a possible genetic phylogeny. A maximum likelihood tree estimate can be found using IQTREE (Nguyen et. al), the resulting tree displayed on the command line using Newick utilities (Junier et. al). Rooting the tree at the midpoint with GoTree (Lemione et. al) allows for a better understanding of a possible chronological evolutionary sequence, although midpoint rooting is not always the most accurate due to variations in evolutionary rate between species. A significant gap in the literature, however, renders this methodology our only option for analysis of a rooted tree, as no evidence has been made available for a potential outgroup.

While it may be valuable to consider the independent gene phylogeny, it is perhaps more relevant to analyze the gene and species reconciliation tree, as being more conducive to exploring the gene family history. Using the aligned fasta tree-file given by IQTREE, execution of NOTUNG (Chen et. al) can work to integrate species and genetic data, creating a reconciled gene-species tree. The resulting tree can be converted into an XML object by RecPhyloXML (Duchemin et. al), and from there fed into thirdkind for the creation of a gene-within-species tree, where duplication and loss events are much more easily visualized. In order to see how well supported these nodes are, IQTREE is rerun using ultrafast bootstrap (Hoang et. al); values of 95% or above are typically considered to have good support.

Finally, we must also work to consider protein domain preservation along the phylogenetic history. Using the unaligned fasta files, InterProScan5 (Jones et. al) is run for sequence analysis, compiling and filtering results to only include those in the PFAM database (Bateman et. al). After slightly rearranging the interproscan output, we can then use the resulting .tsv file to annotate the previously created aligned tree-file on EvolView (Zhang et. al).

The code and data files required to recreate the outlined methodology can be found in the following GitHub repository: https://github.com/Bio312/final-project-L02-megham1ndd

Results

Analysis by BLASTP found 8 homologous genes for XP_001618819.2 among 7 different species *Nematostella vectensis, Dendronephthya gigantea, Hydra vulgaris, Acanthaster planci, Branchiostoma belcheri, Homo sapiens,* and *Mizuhopecten yessoensis*. Results were filtered for e < 10^{-10}. Further analysis by Notung and thirdkind allowed for the creation of a midpoint-rooted gene-species reconciliation tree, outlining specific duplication and loss events based on t_coffee alignment data. A visual representation of gene family evolution within each species can be seen below.

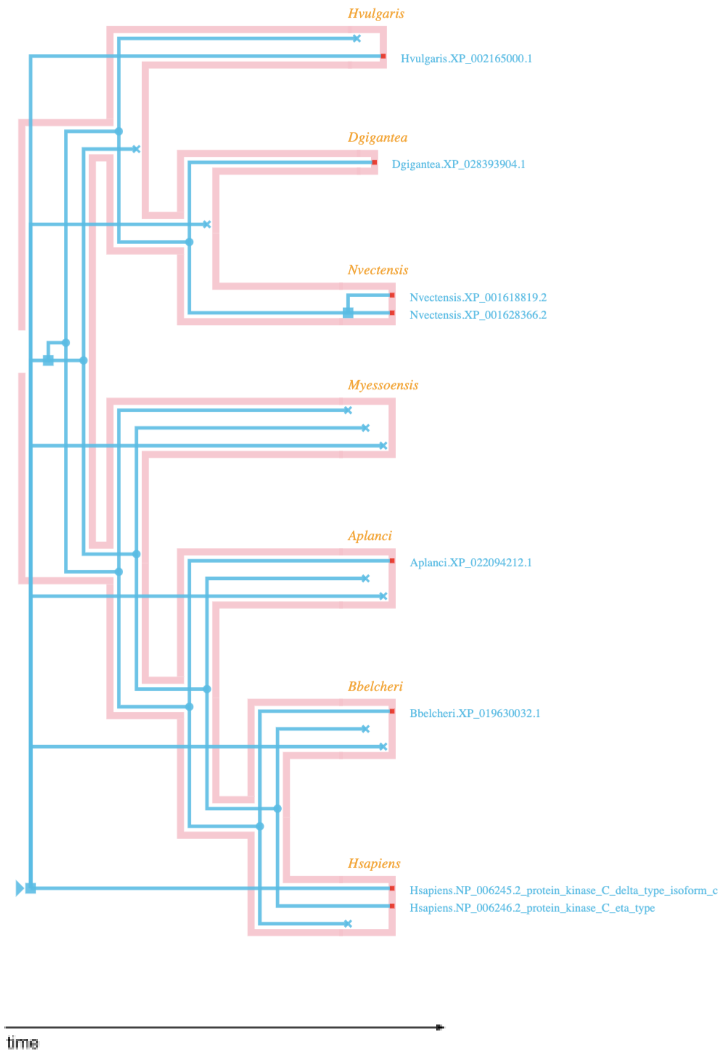

Figure 1. Gene-species reconciliation tree of CRIPak gene and related homologs. Display of gene duplication and gene loss events within respective species. Special superimposition of Notung gene reconciliation tree created using RecPhyloXML and thirdkind.

The reconciliation tree of Figure 1 is primarily characterized by a series of gene losses, in alignment with the lack of homologous genes initially recognized by the BLASTP software. Most species presented in the tree display 1-2 loss events within the species, save for *D. gigantea* and *N. vectensis*. Analysis of *M. yessoensis* reveals only loss events, with no orthologous gene mapping to the original XP_001618819.2 in this species. A single intraspecies duplication event can be found in *N. vectensis*, resulting in paralogs XP_001618819.2 and XP_001628366.2. Total gene duplications and losses are tallied and reported in Table 1.

Taxon	**Duplication events**	**Loss events**
N. vectensis	1	0
D. gigantea	0	0
H. vulgaris	0	1
A. planci	0	2
B. belcheri	0	2
H. sapiens	0	1
M. yessoensis	0	3

Table 1. Table showing the duplication and loss events of CRIPak gene family evolution in selected taxa. Majority of events involve loss of homologous genes in comparative species. Data extracted from the Notung gene reconciliation tree.

Conservation of protein domains are of particular interest in analyzing gene family evolution, considering their relatively slow pace in comparison to sequence evolution. Below is an annotated gene reconciliation tree for the CRIPak family, created via Evolview visualization software and iprscan5 PFAM protein domain annotations.

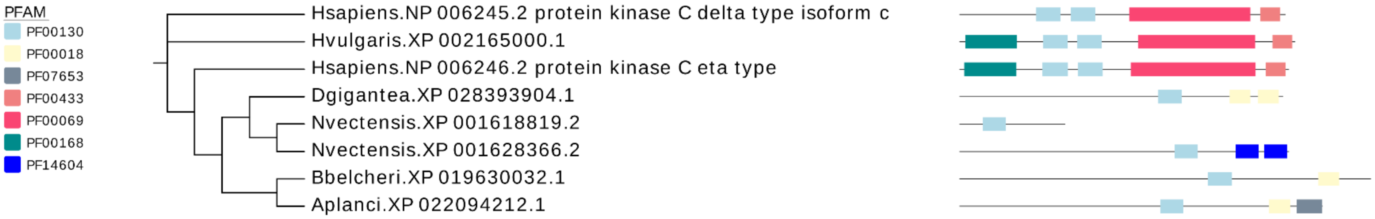
PFAM
PF00130
PF00018
PF07653
PF00433
PF00069
PF00168
PF14604
Hsapiens.NP 006245.2 protein kinase C delta type isoform c
Hvulgaris.XP 002165000.1
Hsapiens.NP 006246.2 protein kinase C eta type
Dgigantea.XP 028393904.1
Nvectensis.XP 001618819.2
Nvectensis.XP 001628366.2
Bbelcheri.XP 019630032.1
Aplanci.XP 022094212.1

Figure 2. Annotated gene reconciliation tree, analyzing protein domain conservation of homologous genes within the CRIPak family. Significant variation is displayed across species, with the exception of a sustained presence of domain PF00130. Created using Evolview software, using IQTREE reconciliation tree and iprscan5 PFAM annotations.

While in Figure 2 there is a substantial degree of variation within the tree, there do appear to be localized areas of domain conservation. Homologs of *H. sapiens* and *H. vulgaris* share domains PF00069, PF00433, and PF00168. Similarly, *D. gigantea* and *A. planci* share the domain

PF00018. Some protein domains, however, are specialized to certain genes, only appearing once in the tree, as PF14604 in XP_001628366.2 of *N. vectensis* and PF07653 in XP_022094212.1 of *A. planci*. Only one domain, PF00130, is shared among all homologs.

Discussion

The evolution of protein domains in this analysis is somewhat defining, especially when considering the high degree of conservation of functional domains in a general biological context. A set of 2-3 distinct groups can be visualized in the Evolview tree, one involving *H. sapiens* and *H. vulgaris*, and another including *D. gigantea*, *N. vectensis*, *B. belcheri*, and *A. planci*, with a possible subgrouping of *N. vectensis*, lacking the PF00018 domain common to the other three species. Interestingly, this functional domain grouping bleeds across the cnidaria/bilateria split. *D. gigantea*, *B. belcheri*, and *A. planci*, all bilaterians, fit into the same group, but apart from *H. sapiens*. Similarly, *H.vulgaris* shares more similarity in protein structure and functionality to *H. sapiens* than to the only other cnidarian involved in the analysis, *N. vectensis*. Consolidating this genetic history with special evolution, we must either assume several instances of convergent evolution, or a significant lack of data lending itself to a reconciliation of an incomplete phylogenetic tree.

The proposed evolution of this gene family relies heavily on a number of loss events. A particularly stringent e value, $e < 10^{-10}$, was used in this analysis, and may very well have overlooked other possible homologous genes in the BLASTP search. Considering the abundance of deletions in the proposed phylogeny, this is more than likely. In fact, no homologs were found in *M. yessoensis*, potentially indicative of a high degree of divergence along the evolutionary timeline. Future analysis may perhaps involve a relaxation of this e value, in attempts to produce a more relevant phylogenetic history.

Unfortunately, due to the lack of available data concerning the CRIPak family and its evolution, there is no good way to confirm the validity of this proposed phylogeny. Very little is published in regards to the CRIPak family, and much of that which is published is in relation to its role in cancer and tumorigenic tissue. The only relevant research involved in the evolution of CRIPak is that published by Watari et. al, involving an evolutionary divergence in a CRIPak-interacting domain of PAK family proteins at the unicellular-metazoan ancestor. Increased regulatory strictness was observed in animal versus non-animal homologs, reflective of an emergent tumorigenic capability in mammalian genes. Further research may thus find it valuable to include more ancient organisms in a similar bioinformatics analysis, perhaps as a means to explore the development of this gene in the context of tumor suppression.

Works Cited

Asmann, Y., Maurer, M., Wang, C. et al. Genetic diversity of newly diagnosed follicular lymphoma. Blood Cancer Journal 4, e256 (2014). https://doi.org/10.1038/bcj.2014.80

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., 1990. Basic local alignment search tool. J. Molecular Biol., 215(3), pp.403-410.

Arvestad, L. (2018). alv: a console-based viewer for molecular sequence alignments. J. Open Source Softw., 3(31): 955. https://doi.org/10.21105/joss.00955

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths‐Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L. and Studholme, D.J., 2004. The Pfam protein families database. Nucleic Acids Res., 32(suppl_1), pp.D138-D141.

Chen, K., Durand, D. and Farach-Colton, M., 2000. NOTUNG: a program for dating gene duplications and optimizing gene family trees. J. Computational Bio.l, 7(3-4), pp.429-447.

Duchemin, W., Gence, G., Arigon Chifolleau, A.M., Arvestad, L., Bansal, M.S., Berry, V., Boussau, B., Chevenet, F., Comte, N., Davín, A.A. and Dessimoz, C., 2018. RecPhyloXML: a format for reconciled gene trees. Bioinformatics, 34(21), pp.3646-3652.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792-1797

Gonzalez-Angulo, A. Lluch, A., Eterovic, A. et. al. ConvertHER: Evolution of genomic alterations from primary to metastatic breast cancer. Proceedings of the Thirty-Seventh Annual CTRC-AACR San Antonio Breast Cancer Symposium (2015). https://doi.org/10.1158/1538-7445.SABCS14-PD3-6

Hoang, D.T., O. Chernomor, A. von Haeseler, B.Q. Minh, L.S. Vinh (2018) UFBoot2: Improving the ultrafast bootstrap approximation. Mol. Biol. Evol., 35:518–522.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G. and Pesseat, S., 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics, 30(9), pp.1236-1240.

Junier, T. & Zdobnov, E.M. (2010). The Newick Utilities: High-throughput Phylogenetic tree Processing in the UNIX Shell. Bioinformatics, 26(13): 1669–1670. https://doi.org/10.1093/bioinformatics/btq243

Lemoine, F. & Wang A. (2017). Gotree. GitHub repository. https://github.com/evolbioinfo/gotree

Nguyen, L.-T., H.A. Schmidt, A. von Haeseler, B.Q. Minh (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies.. Mol. Biol. Evol., 32:268-274.

Notredame, C., Higgins, D.G., Heringa, J. (2000). T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment. J. Mol. Biol., 302:205-217.

Qian, J., Chen, H., Zou, Y., Massion, P. et. al. CRIPAK genomic alterations are associated with indolent lung adenocarcinoma and predicts longer survival. Proceedings of the American Association for Cancer Research Annual Meeting (2019). https://doi.org/10.1158/1538-7445.AM2019-3413

Shen, W., Le, S., Li, Y., Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLOS ONE. https://doi.org/10.1371/journal.pone.0163962

Talukder, A., Meng, Q. & Kumar, R. CRIPak, a novel endogenous Pak1 inhibitor. Oncogene 25, 1311–1319 (2006). https://doi.org/10.1038/sj.onc.1209172

Watari, A., Iwabe, N., Masuda, H. et al. Functional transition of Pak proto-oncogene during early evolution of metazoans. Oncogene 29, 3815–3826 (2010). https://doi.org/10.1038/onc.2010.148

Zhang, H., Gao, S., Lercher, M. J., Hu, S., & Chen, W. H. (2012). EvolView, an online tool for visualizing, annotating and managing phylogenetic trees. Nucleic acids research, 40(Web Server issue), W569–W572.